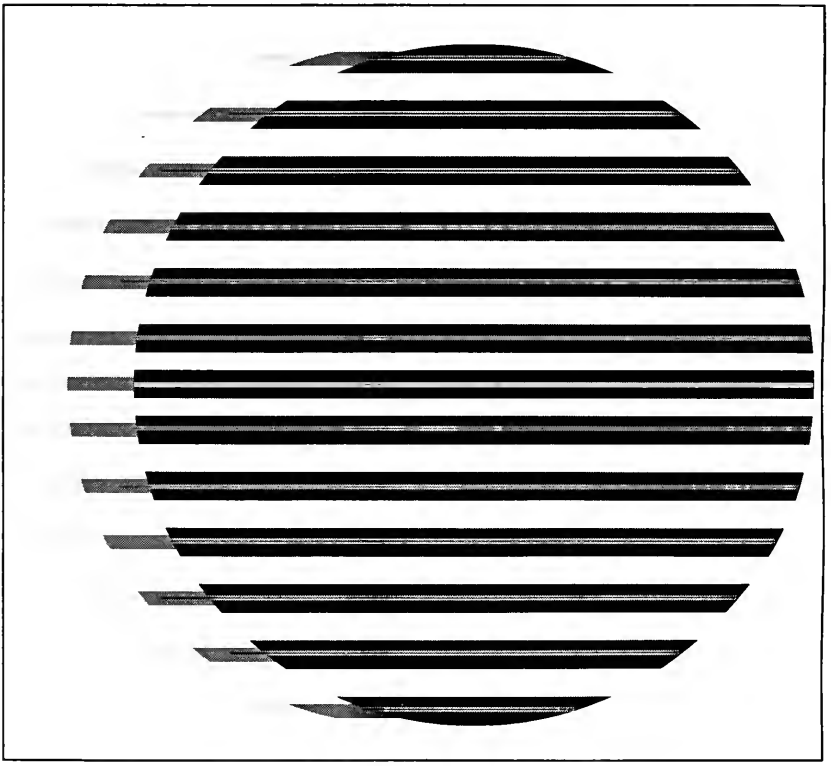
IASSIST

QUARTERLY

VOLUME 21 Summer 1997 NUMBER 2



IASSIST QUARTERLY

The **IASSIST QUARTERLY** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

Information for Authors:

The **QUARTERLY** is published four times per year. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Word processing files may be sent via email to jtstratford@ucdavis.edu. Manuscripts should be sent to Editor: Juri Stratford, Government Information and Maps Department, Shields Library, University of California, 100 North West Quad, Davis, California 95616-5292. Phone: (530) 752-1624.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979*. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

Title: Newsletter - International Association for Social Science Information Service and Technology

ISSN - United States: 0739-1137

CONTENTS

Volume 21 Number 2 Summer 1997

FEATURES

Preservation, Access, and the Multinationals

by Trudy Huskamp Peterson•

Nation states have been the dominant political organizations of the twentieth century. Nation states have national archives. These archives have been dominant, too: developing archival theory and practice, supporting archival organizations, and defining what it means to be an archives and an archivist.

Let us now frame a few research questions that might be posed about the concluding years of this century, the century of the nation state:

- Did the move to invite additional nations to join NATO reflect the notions of identity of populations with each other, with nation state security concerns, or with a desire to lock ever greater portions of the European land mass into one military system?
- Did the recent tumult when Renault announced its plan to shut down its auto plant in Belgium and move manufacturing to Spain's cheaper labor market affect Ford's subsequent decision to continue producing in Germany, even though German firms themselves were fleeing to Central and Eastern Europe?
- Was there a congruence or incongruence between the crumbling status of the Dayton Peace accord in Bosnia and the efforts to rebuild the infrastructure of Bosnia in general and Sarajevo in particular?

To answer the first of these questions, the one on NATO expansion, a researcher will have to have recourse not only to the records of the nation states, but also to the records of international government organizations, NATO in particular, but also the European Union, the United Nations Security Council, and the Organization for Security and Cooperation in Europe.

The second of the questions, that addressing labor movements, popular protest, and industrial activity, would require access to the records of the headquarters of the firms in question, the records of the local subsidiaries (in both the gaining and the losing country), and pan-European manufacturing and labor data from international governmental sources.

The third question, rebuilding Sarajevo in the face of political disintegration, requires access to the records of international philanthropic organizations and other non-governmental organizations, numbering in the dozens of dozens.

Are the questions important? Absolutely. Are records required to answer them? Of course. Are they being preserved? It is difficult to know. What is the likelihood that our future researcher could gain access to this data? In a word: poor.

Let me briefly examine three related questions. First, what are the forces that have made the records of the multinationals significant? Second, what are the factors that make preservation of records a particularly difficult problem with multinationals? And third, what are the current possibilities to gain access to these records?

Galloping Globalization.

Each of the three types of multinationals—international government organizations, international business, and international philanthropic and other nongovernmental organizations—is experiencing galloping globalization. Each affects the other two, directly or indirectly, but each acts autonomously.

The international governmental organizations have had an astonishing growth in the second half of the twentieth century. Unexpectedly, nation states willingly shrank their own powers, agreeing to multinational control structures. Why? Recently a team of researchers consisting of a Russian, a German, and a U.S. economist argued that "the most important national interests of these states [United States, Russia, Japan, and the nations of Europe] converge much more than they conflict. The real interests that the parties share greatly outweigh the interests that divide them."[1] In other words, ceding power has actually been in the national interest.

Be that as it may, the outcome has been the creation of permanent structures, from the European Commission to the World Bank, with a permanent corps of civil servants, unaccountable to any single state, who create records on the most important worldwide issues of our day. While the fears of the anti-UN activists in the United States, who see in the United Nations a conspiracy to establish a world government and extinguish the nation state, are clearly fantasy, it is true that the permanent bureaucratic structures

of the international governmental organizations create the same self-preservation mechanisms that surround any bureaucracy, but importantly absent are the counter-veiling pressures of a citizenry to whom the officials are accountable.

The second type of multinational structure is that of the businesses and commercial establishments. While these have been international for some functions for centuries (think of the Chinese painting porcelain for the European trade), the late twentieth century difference is in the assembly of goods through multiple nations producing components; in the move from international goods or financial markets into international service providers; and from the speed with which information and currency flows. This borderless market, however, still relies on corporate headquarters somewhere on the globe. These headquarters may be the traditional home of the company, where the corporate officers have their offices, or it could be a single officer in a location that gives the most advantageous tax position for the company.

Once again, however, these companies can set their work practices and employment standards without much accountability to anyone, other than to the owners. In the United States, the Clinton Administration has attempted to forge an agreement with a number of the major clothing manufacturers that will cover their operations world wide. The agreement is for a code of conduct, that would prohibit child labor, forced labor, and worker abuse; establishes health-and-safety standards; recognizes the right to join a union; limits working hours to 60 a week "except in extraordinary business circumstances"; and insists that workers be paid at least the legal minimum wage or the prevailing industry wage in every country in which agreements are made.[2] The problem, of course, is monitoring such an agreement. It is a major step, but it is voluntary, it is limited to the manufacturers in one country, and to one industry in that country. The industry is said to be setting up a policing mechanism, but efforts to introduce transparency—including the access to records—in international business operations are Sisyphean tasks.

Privatization—a world-wide trend—also plays a part in the issues surrounding the records of international business. As governments divest themselves of a particular function, the records of that function vanish from the public sphere into the private. From banking to manufacture of weapons, the public track stops at the corporate door. And when the privatized entity is purchased by a foreign corporation (such as the lightbulb maker Tungsram of Hungary purchased by General Electric of the United States), then the policy of retention and access move from that of a national government to that of the foreign parent.

The situation with the major philanthropic and other non-governmental international organizations is different from either the governmental or business model. These organizations, ranging from Greenpeace to the Rockefeller Foundation to IASSIST, are accountable to members or to boards of directors or, even, to heirs of the original donor. The records of the activity may be centralized or dispersed among national chapters; the sources of capital or the number of members may be publicized or a closely held secret; the actions of the board may be publicly reported or may be absolutely secret. What is clear is that these organizations float above or rest lightly within a single nation's structure.

Preserving the Information Base

All three types of international organizations depend heavily on electronic information transfers to accomplish their work. But all three, too, have piles of papers, photographs, videotapes, architectural drawings, and a panoply of other records types. Who decides what to preserve? And, if data can flow electronically, is there any need to move physically other record types—such as paper or video tape or cartographic items—to an archives?

The record of the United Nations and its components on preserving their records is spotty, at best. The central United Nations archives and records service in New York has no authority to control the records policies of the components. Further, with the 1000 person cuts in UN headquarters recently announced, any administrative positions are shaky, especially in something as little valued as the records preservation activity. On the other hand, some UN units have solid records and archives programs, such as the Food and Agriculture Organization or the UN High Commissioner for Refugees. The temporary UN units—such as UNPROFOR—are less likely to have a sufficient records policy to ensure that essential documents are preserved.

Other international governmental bodies, from NATO to the European Union to the World Bank and the International Monetary Fund, are known to have serious records programs. It is not clear that temporary bodies, created for a limited purpose, often as a result of crisis, are adequately documented—just as national governments often have trouble adequately documenting the records of short-term committees, commissions, and boards. Who, for example, is the official secretariat for the documents of the campaign to end female genital mutilation, recently announced as a joint campaign of the World Heath Organization, the UN's Children's' Fund, and the UN Population Fund?

Turning to international business, it is difficult to gain any general picture of the state of preservation of records, given the secrecy that surrounds commercial enterprises. When Royal Dutch Shell, for example, was under attack for continuing to do business in Nigeria, were the records of the Nigerian unit physically transported to the Netherlands? Was the information reported from Nigeria to the headquarters deemed to be sufficient for corporate purposes and the disposition of the records in Nigeria left to chance? Or was it most expedient—if not downright prudent—to destroy the records in Nigeria as soon as possible?

One specific problem for an in-gathering of corporate records in Europe is the policy of the European Union to ensure that if a country wants to keep the records created in its territory, it can. In the case of the Renault controversy, for example, that would mean that Belgium could prohibit the Renault subsidiary from sending its records to the French headquarters, as could Spain. Whether or not this has actually happened, it is possible. In at least one example, France prevented the records of a monastic order from being sent to the headquarters of the order in Rome. Whether any government would think that corporate records are essential for documenting the history of the nation is not clear, but other countries than the European Union—notably Russia—give themselves the legal right to prevent the export of records of a business registered with the government. The best one can say is that at least in such a case the documents would be preserved, although scattered.

The problems with the international NGOs are quite similar to those of the multinational businesses, with the exception that there is less likelihood that they would be destroyed to prevent the release of corporate secrets.

The truth is that, for many records of international non-governmental bodies, whether commercial or philanthropic or pressure groups, there is no logical archival home. If one expects the corporate headquarters to hold records of business world-wide, there would be mass archival storage on the Cayman Islands or in Liechtenstein. In countries where the national archives has a mandate to hold the records of industry or of any organization or establishment within the country, the national archives might be a possible place of deposit. In some countries, however, such as the United States, unless the records of the non-governmental body show the functioning of the government, the national archives does not have the authority to hold them.

Even if the problem of location could be solved, the problem of international transport is daunting. Electronic files can be transported with relative ease (and if they cannot, the matter of carrying diskettes is simple), but operating in many languages, with electronic data recorded in a wide range of fonts, currently represents a major technical problem. Paper and videotapes are another matter entirely. Shell advertises that it operates in 120 countries. McDonald's operates in so many that purchasing power parity can be based on the cost of a Big Mac—with reasonably sound economic predictability. It is simply not realistic to believe that these corporations will move any significant quantity of records around the world—it would not be economic, and for these businesses that has to be the bottom line. Turning to the non-commercial sector, there the funds are usually heavily committed to pursuing the mission of the organization and precious little is willingly spent on administration, for that is just the means to the goal. Unless the information itself has value in pursuing the objective of the non-commercial organization (such as documentation of human rights abuses), it is unlikely to command the resources required for preservation.

Accessing the Record

If the record of a multinational activity is preserved, what is the likelihood that the researcher can gain access to it, in any reasonable time? Again, the answer is discouraging.

The international governmental organizations may have a policy or a procedure, but it is often both arduous and not timely. (One outstanding exception is the Historical Archives of the European Union in Florence, Italy.) The records of the cases at the World Court are closed for 100 years. The International Monetary Fund recently balked at a request for access for official business by an employee of its sister institution the World Bank. Ironically, the closed policy and restrictive conditions of some of the international organizations spill over into the national practice; a recent attempt to adopt a Freedom of Information Act in Latvia, for instance, failed (according to a knowledgeable observer) because Latvia hopes to join NATO and the opponents of the act argued that an open access law would run counter to NATO practices!

Turning to international corporations, the record is opaque. Corporations like Coca Cola have an archives and an access policy; so does the Walt Disney Corporation, some multinational banks, and others. But the records policies of most are unknown.

NGOs also probably present a mixed access picture. Here there is almost no data about the actual access conditions. Let me, instead, give you an example of the archival challenges of the Open Society Archives as an archives of a major international NGO.

The Open Society Archives is the archives for the world-wide network of Soros Foundations. From their headquarters in New York, foundations operate in nearly forty countries world-wide—from Mongolia to South Africa to Estonia to Guatemala. In addition, there are major philanthropic activities in the United States. The Archives itself is located in Budapest, Hungary, which serves as the de facto European headquarters for the Open Society Institute (as the Soros foundation is officially known). In addition to the records of the foundations themselves, the Archives holds by contract the records of the Research Institute of Radio Free Europe/Radio Liberty, in which we find almost every language of Europe and Central Asia; the records of the Index on Censorship, with world-wide languages; and so on.

Obviously, for us fonts and languages are major issues. But so, too, is what to preserve in what format. Because the foundation network is completely networked and dependent upon electronic communications, we are looking hard at the electronic data in the foundations, to see what can be transferred to the Archives electronically and thereby provide a basic portrait of the Foundation's

activities. We also are capturing the electronic traffic broadcast within the foundation network in an innovative electronic storage program: this should give basic outlines, too. We cannot reasonably ship paper or audio-visual records from all points of the globe, yet some records in non-electronic formats are indispensable for providing a picture of what the foundations are achieving. For example, the Romanian Foundation has for a number of years provided grants to companies to take public opinion polls, thereby providing an unbiased source for evaluating attitudes toward issues. The polls are taken by different organizations, and the result is published in a continuing series of hard copy publications. The data is invaluable for research, but to preserve it we have to preserve the hard copy report. Similarly, the videotape of the Roma microlending project by the Hungarian foundation exists only in video format. Books published by foundation grants, documentary films supported by them, and all manner of sound recordings are also part of the legacy.

For us, the mission of the Open Society Archives is to document the foundations and also to preserve information on the period of communism and post-communism in Europe and to document the movements for human rights world-wide. But with the exception of a few other foundations, we are probably a unique philanthropic organization that is willing to consider the historical importance of its records.

What is to be Done?

The issue of preservation and access in archives of intergovernmental organizations has been repeatedly discussed during the last decade. In 1990 at the International Congress of Historical Sciences in Madrid, Charles Kecskemeti, the Secretary General of the International Council on Archives, argued for an archival policy in major intergovernmental systems. This was followed by a paper to the 1995 International Conference of the Round Table on Archives by Liisa Fagerlund, in which she called for "harmonized standards and procedures" and for exploring the possibility of depositing the archives in co-operating archival repositories (such as national archives) "in sites with major concentrations of United Nations system organizations."[3]

In many democratic states, a rock of the social order is the principle that citizens have the right to know what the government is doing and has done—the basis of Freedom of Information legislation. The international monitor group, Freedom House, now estimates that 60% of the governments in the world have a democratic form._ If this is true, then it follows that organizations made up of democratic states should, themselves, have democratic management—in the instant case, a principle of preserving important archives and providing access to them in an open and consistent way. Pressure by member states is critical to ensuring that a discussion of preservation and access goes forward. The goal is democratization of the data.

Unlike the obvious pressure path of citizen to national government to international organization, when we turn to international business we have no such levers. We can assume that business will do whatever it believes good business practice to be, without regard for future research and for history unless it suits the corporate purpose. But we really know very little about the actual situation in the Fortune 500 companies, not to mention such emerging giants as Gazprom or LUKoil. Here I believe the next steps are (1) to survey the actual preservation and access practices in the Fortune 500 companies, giving special attention to the status of the records of offshore subsidiaries; (2) to launch a concerted effort to encourage the companies to use international standards to describe the historic records they do hold and, to the extent possible under corporate guidance, to share that information electronically. Only with some survey results in hand will the research community be able to assess the preservation and access needs.

The records of international philanthropy are somewhere in between the two. If human rights activism uses the politics of guilt, it should be possible to use that same argument with key philanthropies to have them preserve and make available their records. After all, these organizations are or hope to be change agents, and it is important to them to have a means to measure their effectiveness. Records do that.

By suggesting these few activities, I am conscious of the admonition of the Book of Daniel: "Many run to and fro and knowledge shall be increased."[4] I also remember Mao Zedong's wrong-headed idea, as reported in the famous Little Red Book, that "investigation may be likened to the long months of pregnancy, and solving a problem to the day of birth. To investigate a problem is, indeed, to solve it."[5] Survey for the purpose of surveying, preservation for preservation's sake or access for access' sake is not the goal. The goal is, rather, that we take the steps now to ensure that the twenty-first century embarks on the preservation of its history, a history—I am convinced—that will have as a critical component the actions of the multinationals.

[1] Graham Allison, Karl Kaiser, and Sergei Karaganov, "Towards a New Democratic Commonwealth," December 12, 1996 draft, pp. 1-2 (copy in possession of the author).

[2] Dress Code," The Economist, April 19, 1997, pp. 54-55.

[3] Liisa Fagerlund, "Status of records of the United Nations system" (copy in possession of the author).

[4] Daniel 12:4.

[5] As quoted in Paul Theroux, Riding the Iron Rooster: By Train through China, New York: Ivy Books, 1989, p.72.

* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

A Glimpse at the Future of Social Science Statistical Data: new forms of data analysis, new types of access, and new issues for data providers

*by Stephen E. Fienberg**

Abstract

The world of computing and communications is in enormous ferment and statisticians and statistical data users need to pay attention. The Internet has changed forever how we think about the issue of access. At the same time, new forms of data are emerging as alternatives to the traditional numerical responses that survey methodologists have dutifully encoded for use in statistical analyses. Survey data sets of the future may well consist, either through direct collection or forms of record linkage, of combinations of traditional numbers, text, images, sound, and even symbolic summaries. New statistical methods will need to deal with such mixed media, and the new data and methods will raise new issues with regard to such topics as confidentiality. As a consequence, the ways in which organizations provide and individuals access statistical data bases are surely going to change in a radical fashion. This presentation offers some thoughts and speculations on these topics.

Introduction

I am pleased to join you here today. Those involved with social science data archives around the world are providing an extremely important service to highly varied professional communities in support of what I take to be a fundamental principle: the access to research data. There are those who still believe that we should restrict access to data, both government data and those collected by researchers in universities, for confidentiality and other reasons. I believe such thinking is fundamentally flawed, and that the new world of communication and computing will ultimately lead us in the direction of unrestricted access. In fact this is one of the themes of my talk.

The past fifteen years has produced a remarkable revolution in computing and communications, and we are regularly told about the radical changes we can expect to see in the near future. It is that future I plan to comment on this morning, and the likely impact on the way all of us do our work. My message is that "the future is now" and, if we work to take advantage of what it offers, we will change what we do in a radical fashion. My talk consists of brief glimpses at the past and present, and a longer look at the near future. I'll end with some thoughts and speculation on how social science data archives and government statistical agencies can improve the accessibility of the data they collect and produce, and how statistical users should be thinking about the new environments for databases.

The Past

I stand before you as someone whose professional career began in a different era, with the "so-called" mainframe computers of the 1960s, and I remember quite vividly my first exposure to government statistical data bases at the University of Minnesota in the early 1970s. In the building next to us, colleagues had established one of a handful of computer centers nationwide, about seven I believe, devoted solely to the analysis of U.S. census public use tapes. Those of us outside the center would not have dreamed of having direct access to any of these data, except in the summary form released in print by the Bureau. Instead, a full-time center staff responded, after a day or so delay, to requests for summaries or cross tabulations, provided that the data released did not violate what were then the U.S. Census Bureau's confidentiality guidelines.

Statistical agencies have often been in the vanguard of technological change and they have used innovations in computing to change the tasks of their employees and the accessibility of their data. For example, the U.S. Census Bureau employed, some 50 years ago, the first large computer outside of the military. Today, however, changes in computing and communications are occurring at a dizzying pace, and few government agencies can afford to lead let alone follow.

The Present

The world of computing and statistical databases has changed quite dramatically since the early years of computing to which I have been referring, most notably with the rise of the personal computer and the development of distributed computing environments linked through networks that make resources that are thousands of miles away seem as if they are across the campus or down the hall. For example, the results of the 1990 U.S. decennial census are available on compact disc, and anyone with a PC and CD-ROM reader has the physical capacity to carry out innovative analyses of official statistical data, and large chunks of the data are also accessible from the Internet in various forms. And we are slowly catching up with the

missed opportunities of the past. For example, at the University of Minnesota, historians are assembling a WWW site with PUMS (public use microdata files) for all censuses extending back almost to the US Civil War.

Today, the U.S. Census Bureau is, at long last, beginning to think about the possibility of direct World Wide Web (WWW) access beginning with files from the 2000 census of population and housing. But what such unrestricted and unfettered access means for confidentiality may require new considerations. How can data archives and government statistical agencies plan for the uncertain future ahead? At a minimum, they must try to anticipate the types of demands that statistical users will make, and think in terms of leading once again.

The Future of Computer Networking and Collaborative Statistical Research

I now turn to the future, cast largely in terms of activities that I engage in as a statistical researcher. You will need to translate components into a framework more directly relevant to your personal interests. But do not think that what I will describe is a fantasy; each piece of it exists today in a usable or at least a semi-usable form. It is simply that most of us have not yet had occasion to assemble all of the pieces together in a single place to exploit for our own work.

To begin, I describe my environment at Carnegie Mellon University. On my office desk sits a computer workstation with multiple processors; it has far greater capacity than the largest computer available only a decade ago. This workstation is part of a department local area network consisting of about 50 similar workstations and a variety of servers, as well as a graphics lab with several more powerful graphics devices and a terabyte of rapidly accessible memory, all of which are capable of being linked for complex computing and simulation tasks. Our LAN in turn is part of a campus-wide network which has an access point in every classroom and every faculty and staff office on campus as well as in the dormitories and a number of special student computer laboratories. This allows me access to specially programmed and configured machines for language translation and text processing. We link directly to the joint CMU/University of Pittsburgh Supercomputer Center, which is part of the high speed backbone for the U.S. Internet infrastructure. This allows me direct links to resources from businesses, government statistical offices, and universities all over the world. Attached to my computer is a laser printer, a scanner, a compact disk player, stereo speakers and a number of other electronic devices, not all of which are depicted here. My vision is about how statisticians can expect to use this equipment and related technology to facilitate their day-to-day work and collaborations with colleagues around the globe.

It is 9 a.m. on a bright May day, in 1998, and I have just entered my office at Carnegie Mellon University. My workstation is already on and I sign in with my name and password. I open several windows on the screen including my WWW browser, with an applications menu that includes the department's electronic mail system. As I open my electronic mailbox, a message arrives from a colleague, Guido, at the Catholic University of Chile in Santiago. The message includes a draft section for a paper we are working on with a colleague at Statistics Netherlands, Leon and another document whose format I do not recognize. A covering note from Guido describes the additional file, which includes a set of 50 variables which he would like to merge into our data base for estimating the size of the population of several countries using multiple sources. He explains that he acquired the data only yesterday for the U.S., Chile, and five of the EC countries from the web homepages of their statistics agencies, and the datafile consisting of several gigabytes of data was included in the unidentified compressed file I found a few moments ago.

Our paper deals with population size estimation using new statistical techniques for multiple-recapture analysis, and it is based in part on a new probabilistic matching approach which extends the widely-used Fellegi-Sunter method of record linkage. We now have four sources of data for samples from each of the several countries. In his message, Guido explains how he has developed a new way to produce the posterior probabilities of matches for our analysis and that he has also prepared a new program to display the results with dynamic graphics. He suggests that we arrange an interactive video session with Leon in The Netherlands, during which we can experiment with this new program and fine-tune our methods. I send an e-mail message to Leon, who is working late at his office preparing his section of the paper which utilizes text information recorded in survey interviews to supplement responses to a series of questions on race and ethnicity. He responds in minutes suggesting that we begin immediately.

Through customized menus on my WWW browser, I now activate the our current draft manuscript, the results of the data analyses, and the video teleconferencing system. Displayed on each of our screens are live video images of the three participants, in this case Guido, Leon, and myself, and a joint interactive "whiteboard" workspace that we can each manipulate. Guido begins to demonstrate the matching algorithms and his newly developed dynamic graphics tools, but Leon and I occasionally intervene and adjust the procedures and the settings. As we watch the revised program execute we discuss how to alter the text of the paper describing the graphical tools and the data summaries. Leon explains how he proposes to use the text information on race and ethnicity to reclassify respondents by social group, and to correct the probabilistic matching algorithm. We then create a video of the dynamic

graphical display of our probabilistic matching results for The Netherlands. For our method of population estimation, we include all of the extra variables in the database that are not part of the matching algorithm as part of a covariance adjustment for heterogeneity.

The paper we are working on is being submitted to a new electronic statistical journal begun by statisticians at the *University of Stockholm* in collaboration with *Statistics Sweden*. The data base and video have actually become part of the paper we hope will be "published" in the journal.

I then do an electronic library bibliographic search using a local file of statistical titles and key words to locate additional references for the paper with Guido and Leon, and I e-mail them copies of what I find. Carnegie Mellon's campus library is part of an electronic service providing computer access to over 1200 journals, and I can receive faxes of requested articles within 24 hours for those journals that are not electronically archived. I can copy them into my files and share them with my colleagues who do not yet have direct access to this service. Leon in particular is hampered by access, because of the limited technical library at Statistics Netherlands, and the unwise agency policy on restricted access to the WWW.

Before leaving the office I open another web page and locate a weather summary for Odense where I have a meeting the next day. And I move the files from my workstation onto my portable computer to use for my live presentation, as I have done today.

The Future is Today

Who can predict the future with any certainty? As a statistician, I know that such forecasts are fraught with error. In about 1940, the British mathematician G.H. Hardy was asked to comment on the usefulness of the fruits of his research on number theory. He replied that the work had little likelihood of any real world application. This turned out to be incorrect and the past decade has seen a series of major new applications of Hardy's work in number theory. Similarly, there is an apocryphal story about the applied mathematician John von Neumann, who helped to develop the modern computer. At an early stage in this work he suggested that one very large computer is all that would be needed to solve the entire world's computational problems. He would be truly astonished by the new world of computing and communication that we enjoy and take for granted today.

Were I wrong in describing the future of computing and telecommunications to you, I would have placed myself in superlative company by invoking the names of Hardy and von Neumann. But the statistical tools of prediction are not what I have used in preparation for this talk. Rather I have simply told you about things that I have already done myself, some that I have actually seen but not used, and others that I have at least read about. My seemingly futuristic description includes several activities that I actually engaged in over the past month or so, not just contemplated for use on that day in May 1998, when I hope to return to Odense for another meeting. For example, a computer-based video teleconferencing system similar to the one I described exists, but I cannot afford to use it in my collaborations with Guido and Leon. Both of them exist too, by the way, and we collaborate and interact electronically, although not as a trio. Moreover, WWW access to large scale government statistical databases exists but it is limited at best, as is the harmonization of data that would allow comparable analyses across all EC countries. Here are some further caveats to the story I have told:

- The machine translation programs and text processing machines I described in passing was borrowed from my colleague and Dean of Computer Science at Carnegie Mellon University, Raj Reddy, who for several years has been working with others on just such a translation program project.

- The shortcomings of commercial telecommunications systems and the limited channel capacity of several links on the Internet make the kind of real-time quality interactive video/computer teleconferencing I described prohibitively expensive. The transmission of digitized video and audio requires advances in compression technology and alternate modes of transmission. But cable television systems in the U.S. and here in Europe are now providing Internet access to selected areas on a trial basis and in essence the capacity to do the very things I've speculated about.

- Statistics Sweden does not yet sponsor an on-line electronic statistics journal, although it does publish a traditionally-printed journal of very high quality, *The Journal of Official Statistics*. But electronic journals do exist and they are the focal point of the publication strategies for many professional societies, and some have just begun to include videos and interactive programs. The type of on-line access to electronic journals that I described is the goal of many of these groups, and it complements the electronic access to data.

- Most statistical methods still deal solely with numerical data , usually in the form of an $n \times p$ array, or perhaps hierarchically structured. This remains the focus of most data archives as well. But a number of researchers now work with the analysis of images, and others are interested in text data. In fact, one of the goals of a new interdisciplinary Center for Automated Learning and Discovery at Carnegie Mellon university is the development of tools for the analysis of mixed media data, including numerical data, images, text, sound, and symbols, etc. Learning to work with mixed media

presents new challenges to data archives.

• Then there is my own research on multiple-recapture methods (e.g., see Darroch et al., 1993), especially with probabilistic matching (Ding and Fienberg, 1996), but which has yet to progress to the kind of implementation involving text data that I represented. Nonetheless, I predict with some confidence that someone present here today will be working with such statistical tools before the end of the decade, in a fashion not unlike that which I have described.

Making Data Maximally Accessible to Meet Diverse Statistical Needs

As I have tried to suggest, the ways in which we provide and access databases are surely going to change in a radical fashion in the next couple of years. How should the providers and the users be thinking about such changes? My answer is via a new kind of accessibility. For me, accessibility involves a number of different dimensions, and I'll end by addressing three of these briefly: physical access; software reformatting data to meet user specifications; preserving confidentiality of individual responses.

Virtually every statistician and social scientist's desktop has on it a computer with substantial capacity to analyze data using one or more statistical packages that allow for exploratory data analysis as well as more formal statistical methods and graphical diagnostics. This capability is available today to university students and policy analysts in business and in local and state government, not simply specialists in statistics. What is not present for all, but what will be very soon, is the communications and networking capacity I described this morning. Thus I argue that government statisticians now must face the reality of what the users of their data would actually like to have in the way of data access and do in the way of statistical analyses. Statistical agencies can no longer simply release selected cross-classifications or complex files that require the agencies' own computer programs for analysis; they need to facilitate data linkage across data-bases, and provide data in a form suitable for analysis using causal models and prediction equations. And users will demand careful on-line documentation as well as access to virtually complete databases, formatted in ways to facilitate their analysis.

Because of changes in data access, agencies and data archives also need to develop software to function at the interface between on-line files and the kinds of analysis files users require. On the WWW, such interfaces need a transparency not present in current approaches and they will involve programs for formatting, extraction, and even for statistical analyses, especially if specialized analyses are required for proper use of survey data.

Perhaps the most challenging task facing statistical agencies in this new environment is the preservation of confidentiality of respondents. Never before will so many have had access to so much data. In such circumstances, the opportunities for malfeasance will inevitably grow. The simple solutions for data disclosure limitation may no longer be as effective, given an intruder with unlimited access to other databases, on the same or similar respondents and extensive processing capacity for record linkage and matching. The challenge for data archives and government statistical agencies will be to provide "complete" data on samples of respondents, but in a form that makes difficult (although not impossible) the task of an intruder attempting to gain information of specific individuals or enterprises. This requires new ways of thinking about public-use statistical microdata files that are consonant with modern statistical theory - - the focus of much of my current research (e.g., see Fienberg, Steele, and Makov, 1996; Fienberg, 1997). But this is a story for another day.

References

Darroch, J.N., Fienberg, S.E., Glonek, G.F.V., and Junker, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88, 1137-1148.

Ding, Y. and Fienberg, S.E. (1996). Multiple sample estimation of population and census undercount in the presence of matching errors. *Survey Methodology*, 22, 55-64.

Fienberg, S.E., Steele, R.J., and Makov, U.E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and loglinear models. *Proceedings of Bureau of the Census 1996 Annual Research Conference.* U.S. Bureau of the Census, Washington , DC, 87-105,

Fienberg, S.E. (1997). Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research. Background Paper prepared for the Committee on National Statistics, Washington DC.

* Prepared for presentation at the plenary session on "Statistical data Producers," IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science at Carnegie Mellon University, Pittsburgh, PA 15213-3890, U.S.A. He is currently a visiting researcher at Statistics Netherlands.
This work was supported in part through a contract from the U.S. Bureau of the Census with Westat, Inc.

A Digital Library for an Academic and Research Community

by José Luis Borbinha & José Delgado *

Abstract
In the first part of this paper we discuss and define the concept of a networked digital library. We define it both as a new tool for virtual communities engaged in the production and dissemination of information and knowledge, and also as a potential active member of those communities.

In the second part of the paper we present the ARQUITEC project, a trial conceived to assess the defined concept of a networked digital library. ARQUITEC is a work in progress that will result in a prototype of a networked digital library for the Portuguese academic and research community.

Introduction
The actual and future impact of the Internet in our society is one of the most complex and participated discussions of the moment. An emerging issue of this discussion has been the redefinition of the role of libraries, raising the question of what is a "digital library" in a global networked world.

In the first part of the paper we discuss and define the Internet as a communication medium and as a meeting-place, a "new land" of opportunities for the virtual communities. This vision is discussed in opposition to a common vision of the Internet as just a new distribution medium, in the line of the press, the radio or the television.

Based on that discussion, we define the concept of a networked digital library. A networked digital library is seen not only as a repository for data and information, with the traditional missions of preservation and dissemination of knowledge, but also as an active partner with the potential to stimulate, support and register the process of creation of that knowledge.

In the second part of the paper we describe ARQUITEC, a prototype of a networked digital library for the Portuguese academic and research community.

ARQUITEC is a joint effort undertaken by INESC (an R&D institute), the Portuguese National Library and JNICT (the Portuguese R&D funding agency). The purpose of ARQUITEC is to set up a prototype of a networked digital library for the Portuguese research and academic community, which will be used to test the concept and the technology.

The vision
The net isn't 30 million people, it's tens of thousands of overlapping groups ranging from a few people to perhaps a couple of hundred thousand at the largest" (O'Rally, 1996). It has been broadly pointed out that the information technology in general, and the Internet in particular, has been supporting the existence of virtual communities, defined as communities of individuals sharing common interests, but that are not geographically confined. Evident demonstrations of that reality are the existing thousands of News groups and electronic mail lists, dedicated to almost all the cultural, professional and political perspectives.

With that reality, the Internet can be defined as a new virtual space, like a new dimension of the physical and temporal world. It offers a real meeting-place and a multidimensional communication medium, with a social function in the genealogical line of the traditional squares, market places, coffeehouses (see the success of the cybercafes) and the telephone. This is a deeper and vaster view than merely defining it as a simple one-way broadcasting medium, such as the press, the radio or the TV, since in the Internet each one can be an equal player, with the same chances to be active as anyone else.

This vision has been already a field of concrete experiences in scientific and academic communities. It was maybe first identified by Paul Ginsparg and Steven Harnad, that coined expressions like "skywriting", "esoteric publishing" and "pre-print continuum" (Okerson, 1995; Harnad, 1990; Harnad, 1991; Harnad, 1995). Harnad presents an interesting perspective on the evolution of the human communication, with the phases of speech, writing, printing and, now with the Internet, skywriting. Skywriting is defined as both a new medium and a new model of communication, interactive, independent of the space and more suitable with the human cognitive process. This is a scenario favorable to the raising of esoteric virtual communities that, by using the Internet for their natural skywriting and pre-print activities, will be able to work and prosper in the production of their knowledge and memory.

With this reflection we can now complete the view of the Internet as the mean (the "ether") that can allow the library, now converted in the networked digital library, to go and meet the community. Networked digital libraries can be important not only for the geographically defined communities (that have already their traditional communal structures), but even more important for the geographically unbounded communities, where they can play as active members in the process of development and creation of knowledge and memory. Table 1 resumes that vision for the networked digital library paradigm.

In what we call the traditional library, the subject is the book. Its value is "sacred" (otherwise it wouldn't have been purchased) and it is stored "for ever". In this scenario authors decide what to write and when to edit the book, while the librarian decides whether to buy it or not. Finally, the librarians expect the patrons to came to the library and request the book. It was more or less like that until the middle of this century, when the industrial development changed it.

The industrial development reduced printing costs, illiteracy and the physical distances, while at the same time it increased the amount of information produced. It is not possible anymore for an individual to absorb all the knowledge produced by mankind, so it is necessary to specialize. The specialization brought thematic magazines, journals, reports, conference, etc. A new subject emergent from this reality is the "paper", which represents a new type of knowledge. It is not "sacred" anymore, but still formal, being validated by the credibility of an editor or a review committee. This knowledge is not intended to be valid "for ever", but to be discussed during a period of time, refined and, in the end, what survives is then sanctified in books (while the journals and conference proceedings are stored in the basement).

It is difficult for the traditional library to follow the specialization; so the library itself becomes specialized, with the mission to serve specific communities. Usually, those communities control now the library content in their own interest, in the sense of who decides which periodicals to subscribe or what to buy. Quoting Nicholas Negroponte:

> The real value of a network is more related with community than with information. The information super-highway is more than a shortcut to all the books in the Library of the Congress. It is creating a completely new global social tissue" (Negroponte, 1996).

In this scenario the library is requested to perform now a more active role. Since the communities are well identified, it is now possible to anticipate their needs and to provide customized services, such as the notification of new issues, advertisement of new publications, etc.

The scenario changes again with the arriving of the computer. With the desktop publishing tools and WWW, everyone becomes a potential publisher. The process acquires speed, and the subject is the idea. With computer networks, electronic mail and News groups, communities intensify their interactions. To produce fast results, ideas are submitted in pre-prints or presented to discussion as position papers in informal workshops. Ideas that succeed in this process are then published in journals and promoted in formal conferences. What will be the impact of this new reality in the library world?

Using electronic mail and WWW, it is easier for the library to reach the communities and provide new services (such as the announcement of workshops, the arriving of new publications, etc.). By the same reason, it is now easy for the users to interact with the library, not only to access Online Public Access Catalog (OPAC) services but, in an extreme scenario, to contribute also with new kinds of meta-knowledge that can enrich notably the library. Examples of such contributions can be the tuning and completing of thesaurus and catalogue (allowing dynamic and collaborative cataloguing), the attachment of annotations and comments to the stored documents (allowing collaborative refereeing, for example), etc.

Paradigms	**Networked Digital Library**		
	Specialized Library		
	Traditional Library		
Subject	The Book	The Paper	The Idea
Knowledge	Sacred	Formal	Informal
Memory	Persistent	Semi-persistent	Volatile
Actors	Author, Librarian	Community, Editor	Community
Dissemination	Very Slow	Fast / Slow	Very Fast
Library role	Passive	Active	Interactive

Table 1: The library paradigms

After this discussion, we will finish with our vision and a definition for the concept of a networked digital library:

A networked digital library is defined not only as an organized repository of data and information, with the traditional mission of preserving that knowledge,

but as a system with also the mission to stimulate, support and record the process of its creation.

It is now our mission to demonstrate how to turn this vision in reality.

ARQUITEC

ARQUITEC is a trial to test our vision of a networked digital library that will result in a prototype of a networked digital library for the Portuguese academic and research community.

The system will be accessible over the Internet, through a WWW interface, and will provide access to different kinds of technical documents (such as papers, reports, theses, dissertations, etc.), in any field of the knowledge. The architecture of the system is distributed, with each participating institution (universities and R&D organizations) managing its own repository (see figure 2). Based on that infrastructure, the National Library will manage an official repository of digital documents.

We intend to use ARQUITEC both as a technology demonstrator and a pilot system to develop, test and consolidate expertise in three identified issues:

Architectures of distributed digital libraries.

Procedures for management and access to the information, comprising gathering, classification, searching, retrieval and management library procedures.

Innovative services for networked digital libraries, to exploit the potential of interaction between the library and the community brought by open networks, such as the Internet.

Concerning the management of the information, the main problems will be the procedures for the remote submission of documents and their classification and search, as well as the creation and management of the official archive.

The central archive is a repository at the National Library, onto which new documents are automatically copied when they are submitted to the local repositories.

Dealing with documents from different fields of knowledge rises an important issue related with their classification and search. The key problem here is the possible integration of different metadata structures (required by the different contexts and communities) and the use of thesaurus.

We will also explore new services to be provided by the networked digital library, such as a filtering service based on the matching of the user profile and documents classification, an annotation service for documents, a collaborative catalogue and thesaurus, etc.

The library collection

ARQUITEC will provide support for a three-steps workflow in the production of information, comprising:

- **Informal documents**: a class of documents usually called grey literature (such as position papers, drafts, preprints, etc.) often useful only in the short/medium term, since it is expected that they will loose interest or they will give rise to refereed documents.

- **Refereed documents**: such as full electronic journals, papers presented in conferences or published in conventional journals, etc.

- **Formal documents:** theses, dissertations, official reports, electronic books, etc.

The increasing scholarly and scientific activity has resulted in the growth of publications rich in new interdisciplinary perspectives. That kind of contents has been raising serious classification problems for traditional libraries, where collections have been classified with catalogues usually defined by static structures. In order to deal with this dynamic classification problem, our digital library should provide users with an interactive catalog of the documents. As illustrated in figure 1, the catalog will be supported by:

- A document index.

- A multi-context and multi-lingual thesaurus (also interactive).

- The user interactions.

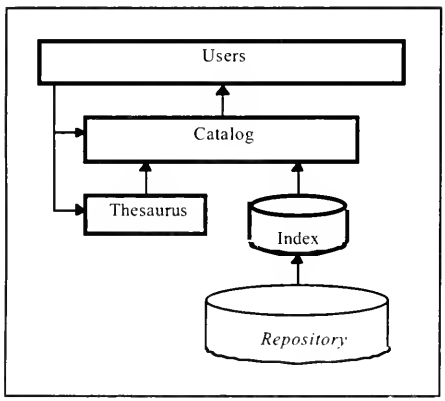


Figure 1: Interactive catalog and thesaurus

The users are able to contribute to the catalog:

- Directly, by suggesting new keywords for documents or questioning existing ones.

- Indirectly, by suggesting new relationships to the thesaurus or questioning existing ones.

For the development and interaction between the catalog and the thesaurus, experimental work was done with MCF (Gutha, 1996), a recent language for meta-content representation. For the thesaurus structure, the ISO-5964 standard was followed (ISO, 1995).

Users and services

Users can access the digital library in one of two modes: anonymous or identified. In order to register a user, the minimum required information is an electronic mail address. However, the users or the system administration can optionally provide other explicit complementary data, useful for some services (such as academic degrees, expertise fields, etc.).

An identified user has a profile, composed of the explicitly provided data and by data implicitly extracted from the history of user interactions with the system. For example, if a user retrieves a document related to a specific subject that is not in their explicit profile, this subject is implicitly added to that user's profile. Pending on explicit confirmation, this new subject will be tagged as a potential interest, which the user can easily change later.

User profiles serve three main purposes:

- Searching: for identified users, the profile is used to rank searching results, highlighting documents that match the profile (but not restricting the access to other documents).

- Filtering: the profile is also used for an information filtering service, supported by electronic mail and by the WWW interface, through which users can be notified, for example, of new documents of potential interest.

- Annotations and catalog tuning: interactive services for document annotation and catalog tuning are also provided. During an interaction with the system, any identified user may contribute also with opinions about document classification, by suggesting new keywords, questioning existing ones or by suggesting changes in thesaurus relationships. These contributions are weighed by explicit parameters of the user's profile (such as the academic degree, for example), and the results of these actions are disseminated by the electronic mailing lists related to the affected documents and subjects. This service gives users a means to interact with the library, not only to access it as an OPAC service but, in an extreme scenario, to contribute also with a new kind of meta-knowledge" that can enrich notably the library.

It is expected that the major part of the documents in the ARQUITEC digital library will be written in Portuguese or English, among other languages. Due to that, the ability to deal with more than one language will be vital for indexing and searching in documents (for example, to recognize common roots in compound words). The success of this task is one of the main targets of our project, having in mind not only ARQUITEC but also its potential application to other similar situations.

A similar problem arises with the diversity of document formats, since we don't impose a unique format. We try to support as many formats as possible, which is nice for the authors but problematic for us.

The integration of such different document formats and languages was done by the development of filters for the indexing and searching modules, rendering the format of documents transparent for the indexing and search tools.

To test solutions for those problems we have been experimenting with publicly available indexing (and searching) tools, such as Glimpse1 and Smart2. These tools have been integrated with Palavroso (Barreiro, 1993) and Correcto (Medeiros, 1995), two successful tools developed by the Natural Language Processing Group at INESC for morphologic and orthographic treatment of the Portuguese language.

Archiving and persistence

A central archive at the Portuguese National Library will be maintained, with a copy of the formal or refereed documents, after copyright has been secured from their producers. This archive will automatically harvest the new documents from the local servers, storing and cataloguing them in a central repository.

A final requirement is name persistence, especially for the documents archived at the National Library.

Depending on whether they are a serial publication or isolated books, printed documents are usually identified by ISSN or ISBN numbers. However, for digital publications such mechanism doesn't exist yet. It is usual to register CD-ROM publications with ISSN or ISBN numbers, specially if they are related to printed publications (such as the CD-ROMs distributed with magazines), but for on-line publications this is not of great help.

The publication of an on-line document is an almost instantaneous process (it requires basically the time to store and to index it in a FTP or HTTP server), and there is no expedient way to require an ISBN or ISSN number for that document compatible with this workflow. Another

important problem raised by on-line publications is that its name, or reference, should not only be an unique reference to identify that object in a specific name space, but should also provide a means to access the document (it must "say" where the object is and how to get it). This is a complex problem, globally known as URI - Uniform Resource Identifier, and its solution has been addressed by the W3C - World Wide Web Consortium3.

At present, the most commonly used form of URI is the URL - Uniform Resource Locator, but URLs have a problem: they are not persistent. If we have a document stored at a server where we need to change the structure of the stored information, the original URL of that document can become invalid, and any reference to it will originate an irritating "Error: the requested document is not valid on this server". In order to prevent that, we must ensure persistent names for stored objects, through some form of URN - Uniform Resource Name.

The problem of naming objects in a digital library was generically addressed in the CSTR project (Anderson, et. al, 1996). That work was reported in the "Kahn/Wilensky Report", from which emerged the concept of handle as an URN (Kahn & Wilensky, 1995). That concept was implemented by OCLC in the PURL - Persistent URL service.

In a few words, the PURL service is based on the existence of a highly reliable server, where it is possible to register pairs of PURLS and related URLs. In its structure, a PURL is a normal URL, with a structure like http://DNS of the PURL server.../object name.... It has a logical meaning that, when used, implies an access to the PURL server that acts as a proxy and automatically translates the logical name to the "physical" URL of the object referred to (a task performed by a simple HTTP redirect).

A PURL service, for all the persistent documents with copies archived at the National Library, will be provided in ARQUITEC. For each persistent document a PURL is automatically and registered at the central PURL server.

The global architecture

Before starting the description of the architecture of our system, we will describe some of the most paradigmatic and related projects already done in the field and whose lessons and results we used for our trial.

Related work

The CORE project started in 1991, and its purpose was to build a database of scanned journals published by the American Chemical Society (Entlich et. al, 1995).

By the end of 1994 they had a database of more than 400,000 pages of full text and graphics (in magnetic tapes and CD-ROM). The text was converted to ASCII and marked-up with SGML (Standard General Markup Language), the database being accessible with dedicated X-Windows interfaces. The other major contributors of this project were the Cornell University, OCLC, Bellcore and Chemical Abstract Service.

The users accepted the results of the CORE project very well, but another conclusion was also that "the task of building and maintaining electronic journal databases remains formidable."

A contemporary and also ambitious initiative was the TULIP project, started in March 1991 and concluded in the end of 1995 (Elsevier, 1996). It was sponsored by Elsevier Science, and involved nine universities in the USA (C.M.U., Cornell, Georgia Institute of Technology, MIT, Univ. of California, Univ. of Michigan, Univ. of Tennessee, Univ. of Washington, and Virginia Polytechnic and State Univ.).

The main goal of the project was to research and test systems for networked delivery and use of scanned journals. Elsevier contributed with the scanned page images, OCR generated text and bibliographic data from 43 engineering and materials science journals. The universities provided solutions to deliver these journals in electronic form to their users. The research focus was on technical issues, user behavior and organizational and economic problems.

When the project TULIP started, the Internet was already a reality, but the Web was still in an embryonic state. Due to that, the delivery technology was based on dedicated graphical clients for X-Windows, MS-Windows and Apple Macintosh, besides alphanumeric clients for mainframe terminals. But soon the maintenance costs were evident, and the project shifted to WWW technology when its advantages and maturity became recognized.

In its final conclusions, the project pointed out that the transition from conventional to digital libraries (defined here as libraries with full digital contents), will take much longer and cost more than commonly thought, mainly due to network bandwidth and storage limitations.

However, and as it was also pointed out by the CORE project, we think that this conclusion can not be dissociated from the approach taken: to scan the original material. For example, it was estimated in TULIP that a typical journal issue, with 20 articles and 200 pages, requires approximately 17 Mbytes of storage, with 16 Mbytes for the scanned pages (in TIFF format). By comparison, the ASCII information resulting from the OCR process requires only 800 Kbytes and the indexing and bibliographic information (in SGML format) requires about 200 Kbytes.

More pragmatic approaches were taken in a series of projects in the Computer Science Reports area. Some of the most representative were UCSTRI - Unified Computer Science Technical Report Index (VanHeyningen, 1994), NTRS - NASA Technical Report Server (Nelson, et. al, 1994), WATERS - Wide Area Technical Report Service (French, et. al, 1995) and CSTR - Computer Science Technical Reports. A common goal of those projects has been easy installation and maintenance of the server sites and support for heterogeneous collections. The idea has been not only to provide scanned versions of printed documents, but also to take advantage of the fact that today it is normal to produce, in the source, those documents already in digital formats (such as ASCII, MS-Word, PDF, HTML, etc.).

In April 1995, WATERS and CSTR projects joined efforts and conceived a new service: NCSTRL - Networked Computer Science Technical Reports Library (Davis, 1995). NCSTRL is a network of servers providing three kinds of services: repository, indexing and user interface. Currently NCSTRL is a worldwide service, with repositories installed in over 60 universities and research centers across the world. NDLTD, a more recent project in the USA, aims to extend that base to provide a generic national digital library of theses and dissertations (Fox, et. al, 1996).

DIENST and NCSTRL

INESC has been experimenting with the NCSTRL technology since middle 1996. We were impressed by its capabilities as a potential framework for future work, especially its open architecture model and its ability to handle documents in several formats. Therefore we decided to use it as the core technology for ARQUITEC. In figure 2 we present the main blocks of that architecture.

The DIENST technology was the main contribution of project CSTR for the NCSTRL initiative (Davis & Lagoze, 1994). The NCSTRL architecture is based on a network of DIENST servers (referred to as S), each one managing a repository of documents (R) the respective index (I) and user interface (UI). The user interface is implemented in HTML, provided through an HTTP server (the DIENST server is written in PERL and its interface to the HTTP server uses CGI). A user can access any server from any user interface, since user searches are always performed in all the indexes.

Optionally, the repositories can be accessed via lite servers (L), the main contribution of project WATERS for NCSTRL. In this case each site only has to provide a metadata description file (M) and have its documents accessible by FTP or HTTP. The lite server converts that metadata to the DIENST format, indexes it, and provides normal DIENST interfaces for the users and for the other DIENST servers. In the specific case of the NCSTRL service, it has only one central server for all the registered lite repositories.

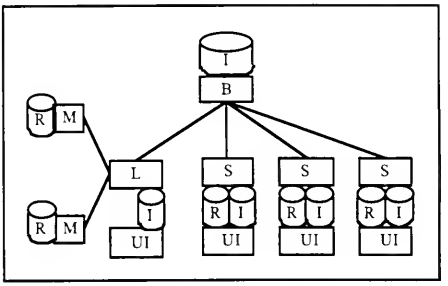


Figure 2: The NCSTRL architecture.

A backup server (B) can maintain a copy of all the indexes, which is useful if one of the servers becomes inaccessible. In that case users will not be able to perform retrievals, but at least they will be able to search and find references to the desired documents.

Finally, our architecture

The architecture of ARQUITEC is distributed, with local nodes managing the local repositories at the universities and research institutes, but all the collections are freely accessible for search from any node. The core of ARQUITEC is based on a modified and extended version of DIENST 4.0. The required modifications occurred at the three modules of NCSTRL, corresponding to three different tasks of ARQUITEC: replacement of the indexing and searching tool, modification of the repository management and modification of the interface.

The original DIENST indexing and search tool had to be replaced by a more powerful catalog, as described. The new requirements implied modifications at the NCSTRL repository interface level, in order to perform full text indexing of as many document formats as possible (such as ASCII, Postscript, MS-Word, etc.), as well as in different languages.

Concerning the management of information, the main generic problems were the procedures for submission of the documents, their classification and search, as well as the creation and management of the central archive.

The submission of documents can be done remotely, with the user authenticated by username and password (stronger security and authentication issues, for which we recognize the importance, were not addressed for now). The submission process starts by the filling and submission of

registration forms, by WWW. Users will be required to provide the location of the original document at an FTP or HTTP server. After that a confirmation procedure takes place: an electronic mail message is sent to the user and the system waits for a reply. After successful confirmation, the document is then retrieved, registered and added to the catalog.

The core of the NCSTRL system was also modified in order to allow the automatic management of the official central archive. In practice this means that the central host, at the National Library, automatically gathers all new persistent indexed documents into a central repository. That repository is used as an official archive, which is especially important for theses and dissertations. It also serves as a mirror repository to provide global fault tolerance.

The NCSTRL user interface was modified in order to support all the described requirements, new functions and services. The modifications were done essentially in the submission of documents (that can now be done remotely), as also in the support of the search task. All the interface components were redesigned to support multi-lingual access (Portuguese and English in the first release).

Finally, a directory for the registered users was added to the system. It is a distributed directory based in the X.500 model, with an LDAP interface (Yeong et. al, 1995).

Future work and open issues

Medium term work will be concerned with the integration of other spaces, accessible by new interfaces at lite DIENST servers. Examples will be interfaces for Z39.50 servers4, useful for the integration of OPAC systems such as the catalogs of conventional libraries, and HARVEST5 brokers, useful for the support of informal publications and other similar material such as mailing lists, source code, etc.

Examples of other identified research issues requiring our attention in the medium/long term are:

- Document structuring: research will be done on using SGML and other alternative solutions for structuring the information objects (a specially interesting issue to be applied not only for the original documents but also to represent the associated annotations);

- Natural language: trials will be done in the classification and search of documents with natural language techniques, with a special concern for the Portuguese language;

- Authentication and certification authorities: the requirements for authentication and certification authorities, for both the documents and users, will be addressed in medium term;

- Legal issues: among generic problems, such as how to assign and observe other properties of the documents (such as terms and conditions and other copyright problems), examples of new open interesting problems in this field are the legal implications of the new objects, composed by an original document and a list of annotations (or just the legal implications of an annotation);

- Long term preservation: how will the official repository survive the evolution of the hardware and software, such as storage technology, operating systems, document formats, viewers, etc.?

References

Anderson, G.; Lasher, R.; Reich, V. (1996). The Computer Science Technical Report (CS-TR) Project: A Pioneering Digital Library Project Viewed from a Library Perspective. The Public-Access Computer Systems Review 7, No 2, 1996. Available at http://info.lib.uh.edu/pr/v7/n2/ande7n2.html

Barreiro, A.; Pereira, M., J.; Santos, D. (1993). Linguistic options and criteria in the development of Palavroso, a computational system for the morphological description of Portuguese (in Portuguese). INESC Report No. RT/54-93, December 1993.

Davis, J. R. (1995). Creating a Networked Computer Science Technical Report Library. D-Lib Magazine, September 1995. Available at http://www.dlib.org/dlib/september95/09davis.html

Davis, J. R.; Lagoze, C. (1994). A protocol and server for a distributed digital technical report library. Technical Report TR94-1418, Computer Science Department, Cornell University, 1994.

Elsevier Science (1996). TULIP Final Report. Elsevier Science Edition. Available at http://www.elsevier.nl/locate/tulip.

Entlich, R.; Garson, L.; Lesk, M.; Normore, L.; Olsen, J.; Weibel, S. (1995). Making a Digital Library: The Chemistry Online Retrieval Experiment. Communications of the ACM, April 1995, Vol. 38, No. 4, 54.

Fox, E. A.; Eaton, J. L.; McMillan, G.; Kipp, N. A.; Weiss, L.; Arce, E.; Guyer, S. (1996). National Digital Library of Theses and Dissertations. D-Lib Magazine, September 1996. Available at http://www.dlib.org/dlib/september96/theses/09fox.html

French, J. C.; Fox, E. A.; Maly, K. (1995). Wide Area Technical Report Service: Technical Reports Online. Communications of the ACM, April 1995, Vol. 38, No. 4, 45.

Gutha, R. V. (1996). Meta-Content Format. Apple Computer. Available at http://mcf.research.apple.com/hs/mcf.html

Harnad, S. (1990). Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry. Psychological Science 1, 342 - 343.

Harnad, S. (1991). Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge. Public-Access Computer Systems Review 2, No 1, 39-53. Available at http://info.lib.uh.edu/pr/v2/n1/harnad.2n1.

Harnad, S. (1995). The PostGutemberg Galaxy: How to Get There form Here. The Information Society 11(4), 285-291.

ISO - International Organization for Standardization (1995). ISO-5964: Documentation Guidelines for the establishment and development of multilingual thesaurus. Geneva, 1985.

Kahn, R.; Wilensky, R. (1995). A Framework for a Distributed Digital Object Services. Available at http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html

Medeiros, J., C.,(1995). Processamento Morfológico e Correccao Ortográfica do Português. Master Thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa, Lisboa.

Negroponte, N. (1996). Ser Digital. Editorial Caminho (Portuguese Edition of the original title "Being Digital", 1995).

Nelson, M. L.; Gottlich, G. L.; Bianco, D. J.; Paulson, S. P.; Binkley, R. L.; Kellog, Y. D.; Beaumont, C. J.; Schmunk, R. B.; Kurtz, M. J.; Accomazzi, A.; Syed, O. (1994). The NASA Technical Report Server. Internet Research: Electronic Network Applications and Policy, Vol. 5, No 2, 25-36.

O'Reilly, T. (1996). Publishing Models for Internet Commerce. Communications of the ACM, June 1996, Vol. 39, No 6, 79-86.

Okerson, A. S.; O'Donnell, J. (1995). Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing. Association of Research Libraries, June 1995.

VanHeyningen, M. (1994). The Unified Computer Science Technical Report Index: Lessons in Indexing Diverse Resources. Second International World Wide Web Conference, WWW'94 Oct. 94, 535-543.

Yeong, W.; Howes, T.; Kille, S. (1995). RFC 1777: Lightweight Directory Access Protocol. IETF Network Working Group. Available at http://www.umich.edu/~rsug/ldap/doc/rfc/rfc1777.txt.

1 http://glimpse.cs.arizona.edu

2 ftp://ftp.cs.cornell.edu/pub/smart

3 http://www.w3.org/WWW/Addressing/Addressing.html

4 http://lcweb.loc.gov/z3950/agency

5 http://harvest.transarc.com

* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

José Luis Borbinha (Jose.Borbinha@inesc.pt) IST - Technical Superior Institute (Lisbon Technical University) Department of Electrical and Computers Engineering
José Delgado (Jose.Delgado@inesc.pt) INESC - Institute for Systems and Computer Engineering Telematics Systems and Services Group

Hyperlinked Eurotrends

by Lorenz Gräf

Abstract
Hyperlinked Eurotrends are a new groundbreaking approach to present codebook data on the World Wide Web. The author created the so-called Eurobarometer Trends Codebook System (EUTRECS). Nearly all available material about Eurobarometers is presented in a clear and userfriendly way. EUTRECS gives fast and easy access to all metadata and datasets. It is the first comprehensive attempt to use internet technology to serve the most basic needs of secondary analysis researchers. EUTRECS is based on four principles: 1) Hyperlinks connect all interrelated variables; 2) A comprehensive search engine gives access to full text retrieval in all Eurobarometer Codebooks; 3) An index of over 1,000 keywords and a classification scheme of over 400 trend variables present an easy way to browse through Eurobarometer questions; 4) datasets and codebooks are available for immediate download. It will greatly enhance the dataservice of the archives. This paper describes the basic features of this system.

Scenario: What a researcher could think
Suppose you are a social scientist and you want to know which questions were asked in the Eurobarometer surveys. You have access to the Internet and you are looking for a database containing question wording of Eurobarometer questionnaires. After having visited all WWW search engines in vain, a friend of yours gives you a hint. You finally get through and a welcome page invites you: "Please type in your query". But how can you know what you are looking for, if you don't know what is in those Eurobarometer studies.

Let's assume, you know already what you are looking for. You are familiar with the topics of the Eurobarometer, but you have never worked directly with Eurobarometer data. Now you want to do some analysis about attitudes towards the common currency. You go to that question database, ask for "currency" and find the items you are looking for. But you are still without data. If your institution is a member of ICPSR, you are lucky and get the desired data quick via Internet. But otherwise you will have to wait until your request for data has reached your home archive and has been processed there. However, you need the data just at this very moment. It would be better for you if you knew someone who has these data already. Send him an e-mail and you will have your data within short time.

Let's think further. You have got the data and have run an analysis. But you are puzzled by your findings. You rerun the analysis, the findings remains the same. After a while you will be wondering if there is anything wrong with that question: "What was the exact wording of question Q34_a in EB 40.1?" Just two days ago you had received the printed codebook . But unfortunately it's in your office and you are on weekend some hundred miles away. So if you are lucky and you know a friend who has specialized in European politics and analysis of Eurobarometers, go to the telephone and give him a call. But otherwise?

Let's think positively. You have taken the codebook with you and resolved your question wording problem. Now you have a new idea: it could be very interesting to compare the attitudes towards the common currency between various professions and diverse levels of occupational prestige. You find a variable concerning "occupation" but you have to construct a prestige scale for that variable. After a few minutes you remember that there was an article concerning this issue some months ago. But who was the author and in which journal has it been published? It would be great if you could use his work and not having to redo the construction work for a second time.

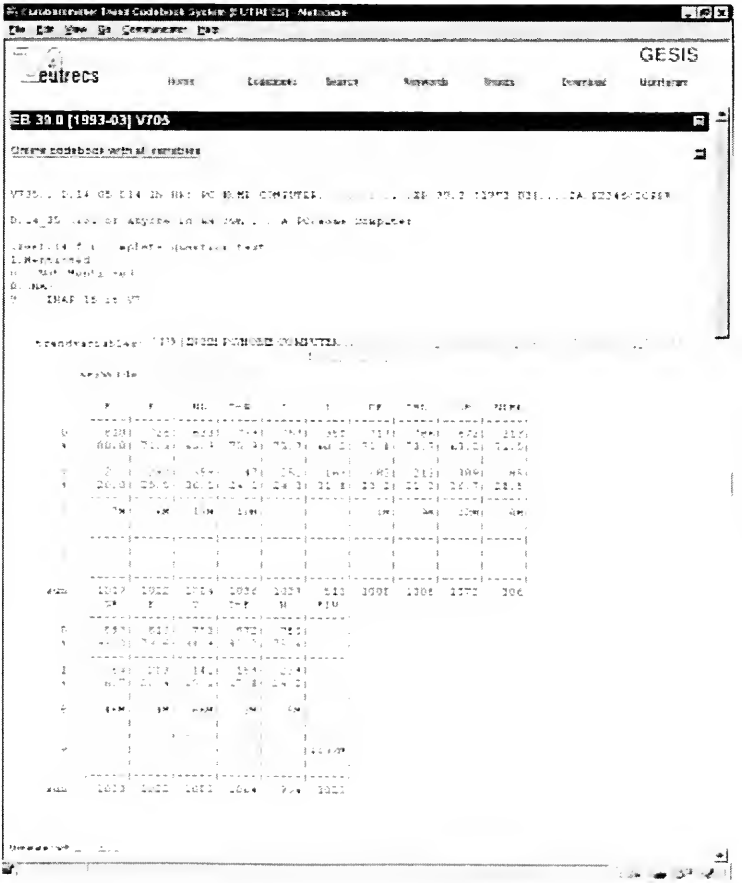
The EUTRECS answer
The Eurobarometer Trend Codebook System gives a solution for all these problems. Those who are not familiar with Eurobarometers can browse in a keyword and trend index. They first scroll through subjects or concepts. If they find something interesting, they click on a link to that variable and have question text and marginals right in front them. Is that question asked more often than once they simply click on another link to follow that item over time. Browsing through codebooks that way, you may find an interesting keyword and you are wondering what else has been asked concerning that topic. So you follow the keyword link of that concept, finding yourself in the keyword index and having a list of questions with similar content in front of you.

Those researchers who are familiar with Eurobarometer studies find a fully searchable codebook database. Every single word contained in a codebook can be found. All the internal linking between similar variables has been preserved. From within EUTRECS users can download every dataset except those which are currently under embargo. This service is open only for non-commercial use and academic purpose. You have to register and accept online the terms of use agreement. Once registered researchers get an username and a password. After having selected one or more datasets they get emailed a transaction id. Within minutes you can have the data. The hyperlinked Eurobarometer web pages provides space for user communication. In the moment there is a list of nearly 800 working papers all about Eurobarometers. If users contribute to that pages there will be much more content in the next future.

EUTRECS in detail

Overview

Eurobarometer are surveys conducted on behalf of the European Commission in the member countries of the EC. They focus on topics concerning the EU. Normally they are conducted in spring and fall every year. Sample size in each country is about 1,000 persons. Data are gathered using face to face interviews. Besides this standard series, there are several so-called flash Eurobarometers. They are conducted by telephone, on a smaller sample size (about 500) and focus mainly on one topic. Since the breakdown of the former soviet imperium every year a survey similar to the standard Eurobarometer is carried out in the Central and Eastern European countries. Two years after the fieldwork the data is normally available for academic researchers. Those who want the data earlier need a special permission of the European Commission. The series of Eurobarometer surveys started in the early seventies. Up to now, there have been conducted more than 70 standard, 50 flash surveys and 10 Central and Eastern European surveys.

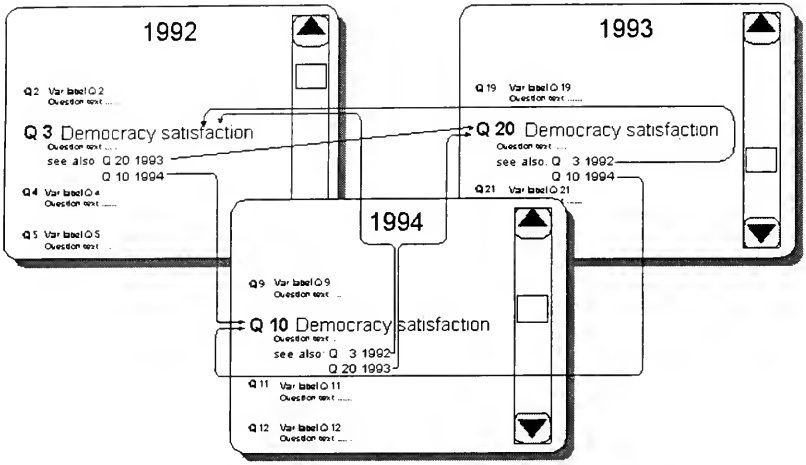


The Eurobarometer site built for the Central Archive (ZA) is based upon the codebooks created by ICPSR, SSD and ZA itself. So far the codebooks are only produced for standard surveys which do not underlie the embargo restrictions. Background information about survey characteristics and topics covered is available for all Eurobarometer surveys. Datasets of all type of Eurobarometer surveys (standard Eurobarometer, flash Eurobarometer and Central and Eastern Eurobarometer) can be downloaded, but this service is restricted to registered users. EUTRECS consist of 32,000 HTML-pages, comprises 1,000 keywords and over 400 trend variables. It is built on basic usability principles and aims to encompass user needs.

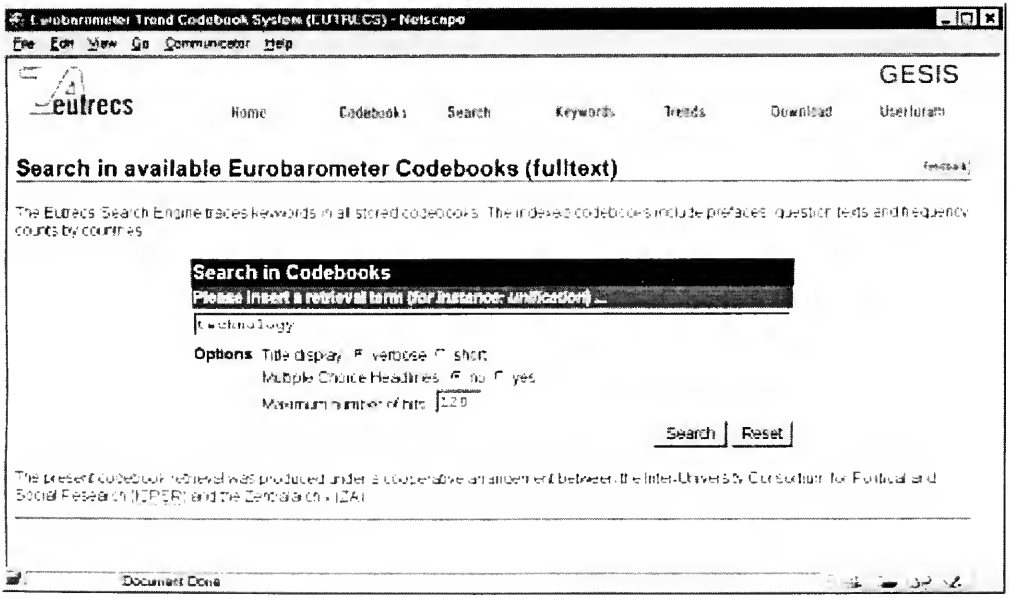
Hyperlinked structure

The Eurobarometer codebooks

make use of the core web technology. Hyperlinks between variables make it easy for the user to explore survey content and to follow his associations meanwhile browsing through the hyper codebook. What does this mean? Suppose you are



interested in attitudes towards *technology*. You look for the item *technology*. Some of the items you find deal also with *computer*. In EUTRECS you can follow the link underneath the item *computer* and find a list of all variables containing this term. By browsing these variables you discover a variable containing information about the diffusion of computer in diverse European countries. You can see that at the beginning of the nineties in the Netherlands and Great-Britain computers were

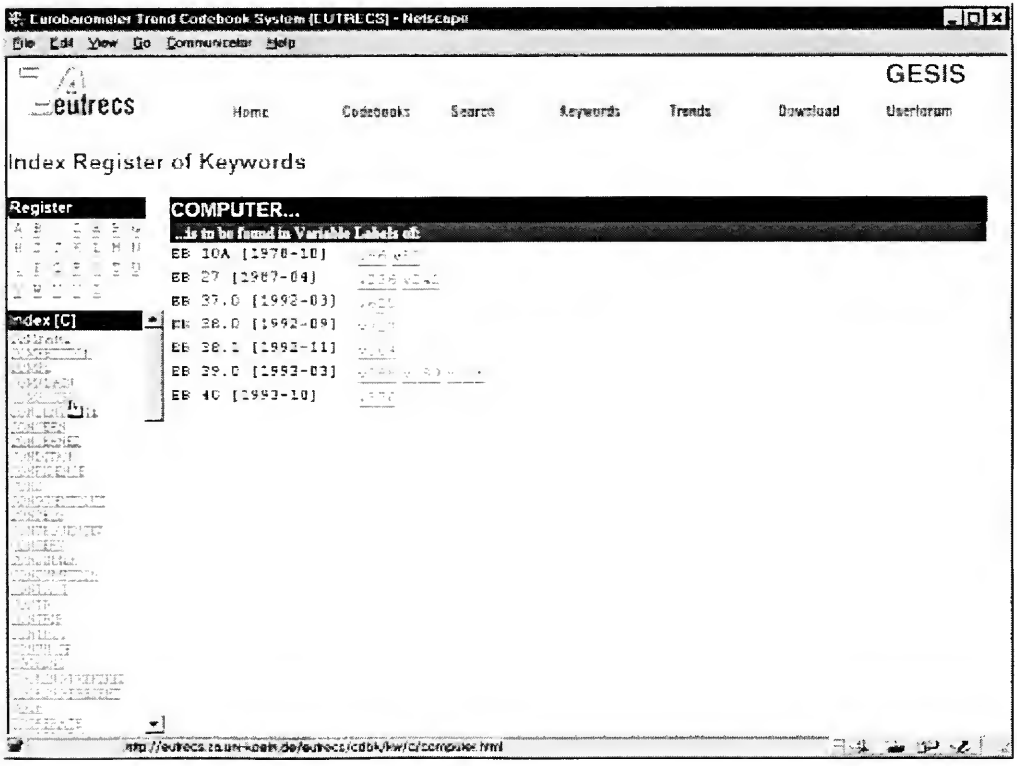


present in a third of all households. At the same time in Germany and France only every fifth household possessed a computer. Now you are interested to know how these figures have changed over time. In other information systems on the web you have to go back to your query result and click on the next item. So if you want to follow a fairly large trend you are forced to go forward and backward some nasty long time. EUTRECS shows on every variable output all other variables which contain the same trend. Therefore it is easy to follow the same variable over time by only clicking on the variable name.

EUTRECS is based on the already available ASCII codebooks. When creating HTML pages EUTRECS needs a continuity table to link identical or nearly identical questions. This table is built in two ways. First, variables with identical labels will be identified and inserted in the table. The second way consists in preparing this table outside of EUTRECS and using the internal codebook information to check the applicability of the pre-given information. As shown in the next figure EUTRECS enlarges the ASCII codebooks and adds links.

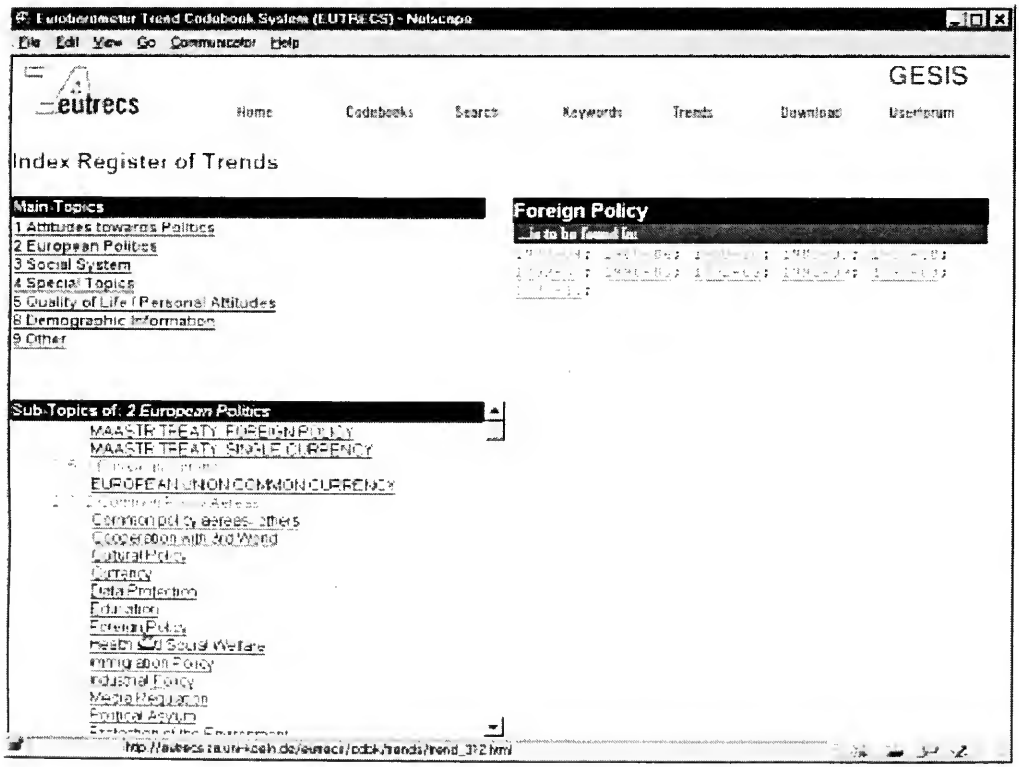
Fulltext search engine
EUTRECS has a built-in search engine based on Freewais-SF. All relevant codebook content can be searched using a fully featured search engine. It is the same engine that serves the CESSDA database. What is unique to EUTRECS is that it preserves all the internal links during processing in the search engine. So, if a user puts a query to locate all variables containing the term *'technology'* on the display of results he finds links to similar concepts, links to other variables of the same trend and a link to the complete online codebook enabling him to explore the context of this variable in the original questionnaire.

Keyword and trend index A web based information system should present as much information as possible in html files. Following this way, researchers can use the web site like a book. In EUTRECS browsing in codebooks is possible.



Comparing EUTRECS to books, the single online codebooks are the chapter of the book. Like in books, identical or similar concepts are dealt with in different chapters. To facilitate navigation, EUTRECS provides not only a free text search engine but also a keyword and a trend index. The keyword index groups variables with similar subjects together. The trend index combines identical or nearly identical questions in different surveys.
Keywords are extracted from variable labels. Doing so EUTRECS can take advantage of a quasi-controlled vocabulary and is not bound to the question wording. But it heavily depends on the quality of the variable label wording. In Eurobarometers this labeling should be improved. Only codebooks produced in the last years contain really good labels but standardization is already on the way. It would be better if we had real descriptors for each variable. But for the time being we have to be content with the material as it is. From the variable labels of the available codebooks we extracted over 8,000 tokens. As a first step1 we implemented a stop-word list and ended up with about 1.000 different keywords. EUTRECS present the keywords in alphabetical order entirely on one screen. You can see in the following figure how easily and fast navigation is. With only three clicks the user gets to the desired variable. A comprehensive list of all keywords is available but it is nearly 2 MB big.

A series of thematically similar surveys offers great opportunities for time oriented secondary analysis. To foster such analysis researchers ought to know which question was asked more than once and in which surveys. The EUTRECS trend index gives the most comprehensive trend register of the Eurobarometer surveys. It contains trend information identified at ZEUS (Mannheim), and at ZA (Köln). But unlike ZEUS we defined trends strictly on a single variable level. In our definition a trend is every single variable that can be traced over time. It is this data that can be compared between different points in time. Now what is the difference? At ZEUS trends are identified, named and differentiated on a concept level. In EUTRECS only questions with identical wording or identical subject have got the same name. Let's give an example. ZEUS presents each item of the question "Which of the following aereas of policy do you think should be decided by the (NATIONAL)



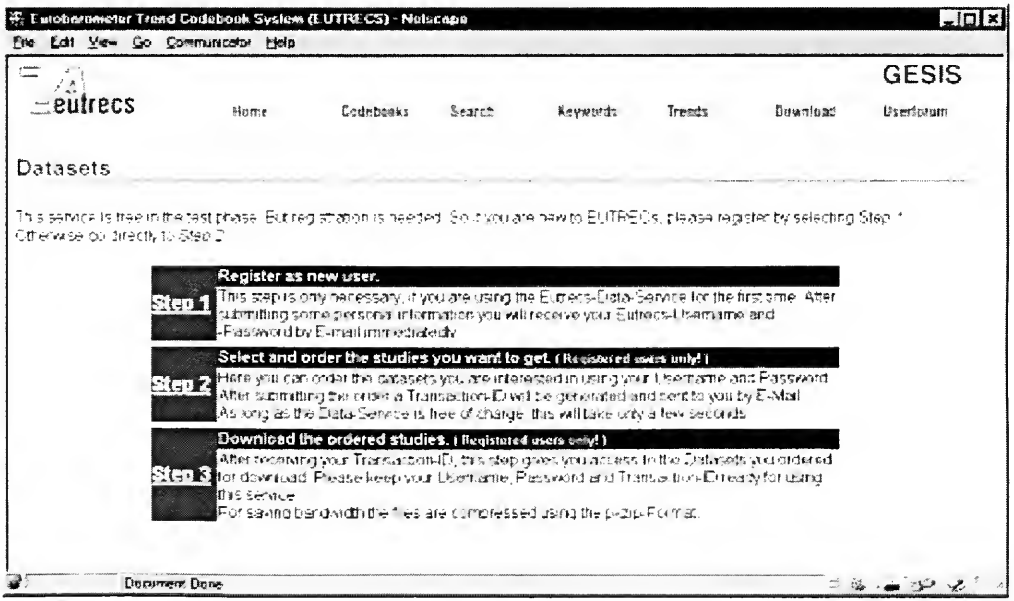
government, and which should be decided jointly within the European Community?"2 under the label 'Common policy aereas'. Whereas in EUTRECS we name the concept 'Common policy aereas' but give the trends the name of the stimulus object, i.e. 'foreign policy', 'currency' or 'education'. By doing so we can provide links between all variables which have the same question wording. Clicking on these links will guide the user through all instances of the trend. As EUTRECS is built out of machine readable codebooks together with each question text full marginal information is displayed. In this version we were not strict on question wording and did not differentiate questions with minor deviations in question wordings. We leave it up to the researcher to decide whether questions were identical or not.

Until now we have identified over 400 single trend variables. To facilitate finding topics we classified the trends in a new classification scheme. In our opinion it was the best method to group similar topics together and avoid getting lost in a mere alphabetical ordering. This classification scheme includes three levels. The third level describes the concept level and is comparable to ZEUS trend names. Beneath that level we have classified the variable trends which we could identify. So we can distinguish between the concept and a measurement level. In the following figure you can see how trends are presented in EUTRECS.

Download
Within EUTRECS the direct download of SPSS datasets and of codebook material is possible. Bandwidth on the Internet is normally small. Therefore we divided the big codebooks in smaller parts of about ten or fifteen variables. But for printing purposes many users want to have entire codebooks. For offline use we give access to all codebooks in html- and postscript format. Keyword and trend indices can be accessed in one-file lists.

Access to datasets is given via an entirely online procedure. We tried to find the easiest and fastest way to deliver data over the Internet without having to compromise to the vital interests of the archives. So we ended up with a three step procedure. In the first step researchers register at the ZA in Cologne. All they need is a functional e-mail address. After registration they get an account and a password via e-mail. Provided with username and password in step 2 they can select datasets. After having accepted the *terms of use agreement* and having paid a fee (depending on the usage conditions of the archive) they get a transaction id by e-mail. With this id they go directly to the download page and get their material immediately.

User forum
The web site we intend to create should be as interactive as possible. It should be the most efficient site for all researchers



doing secondary research with Eurobarometer surveys. We pursue three aims: First, this site should be a platform for the communication between users and the archive and between users themselves. It would be nice, if the archives use this forum to announce any news concerning Eurobarometers (bug reports, announcement of availability and declarations about archive policy). Second, we try to stimulate a shared knowledge forum. In this forum every researcher should announce his findings in the Eurobarometer data. The site now contains a list of working papers done with Eurobarometer material. We would appreciate if researchers uploaded their newest paper for the communication with the scientific community. And we encourage researchers to share their operationalisations and measurement attempts of the data. It would be a really common good to have samples of SPSS or SAS statements there for often needed recoding of Eurobarometer variables. Third, we will give EUTRECS users the necessary software tools to use materials of this site, like ghostview to read ps-files.

What comes next

EUTRECS is going to be developed in a multistage process. In stage one it was important to explore possibilities to present large and complex survey metadata on the web. In the next step we need the help and feed-back of the users to add consistency to the pages. We have to find and correct miscoding, misleading labeling and false classification. It has to be tested how usable the site is and how navigation could be improved. Also the gap in the codebook production should be closed as soon as possible. In the last step we want to incorporate in the data itself trend information and other measurement suggestions of the research community.

1 In future versions a better suited dictionary will be used to improve building of keywords.

2 Exact question wording (EB33 Q30): "Some people believe that certain areas of policy should be decided by the (NATIONAL) government, while other areas of policy should be decided jointly within the European Community. Which of the following areas of policy do you think should be decided by the (NATIONAL) government, and which should be decided jointly within the European Community?"

- Foreign policy towards countries outside the European Community
- Education

WWW: http://infohttpsoc.uni-koeln.de/graef http://solix.wiso.uni-koeln.de/~graef/

* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

Lorenz Gräf, University of Cologne, E-Mail: lorenz.graef@uni-koeln.de

ROADS to Metadata

*by Debra Hiom**

Abstract
ROADS (Resource Organisation and Discovery in Subject-based Services) is a UK Higher Education funded project to design and implement a user oriented resource discovery system. The project is investigating the creation, collection and distribution of resource descriptions to provide a transparent means of searching for, and using resources on the Internet. The system is being piloted on a number of Internet subject gateways, namely ADAM (Art, Design, Architecture and Media), Biz/ed (Business Education on the Internet), IHR-Info (Institute of Historical Research), OMNI (Organising Medical Networked Information) and SOSIG (Social Science Information Gateway). The paper will discuss the background to the project, the type of metadata being collected by the subject-based gateways and the possibilities of cross searching distributed databases, with specific references to SOSIG.

The project uses a standard template for recording information about resources (this was originally based on the Internet Anonymous FTP Archive (IAFA) Template) which is a simple text based record using attribute-value pairs. The simplicity of the ROADS format also provides possibilities of mapping to and exchanging data with other metadata formats, for example the Dublin Core or other standards.

One of the aims of the subject based gateways is to encourage information providers to become involved in the creation of records about their own data in order to make their information as useful and accessible as possible; an approach to this will be discussed.

Background
ROADS[1] (Resource Organisation and Discovery in Subject-based Services) is a collaborative project funded by the Electronic Libraries Programme[2] (eLib) in the UK, to design and implement a user oriented resource discovery system. The ROADS partners are:

- ILRT (Institute for Learning and Research Technology) at the University of Bristol - responsible for user liaison and project management
- Loughborough University (Department of Computer Science) - responsible for the software development
- UKOLN (Office of Library and Information Networking) at the University of Bath - responsible for co-ordinating metadata requirements and issues

ROADS has been created to provide a set of software tools and standards for building and maintaining catalogues of Internet resources. The system allows resources to be catalogued and indexed and provides a searchable and browsable interface to the resource descriptions. The ROADS system is primarily being piloted on a number of eLib funded subject information gateways (under the Access to Network Resources (ANR) programme) who feed into the development of the software. The gateways using ROADS are:

- ADAM (Art, Design, Architecture and Media)[3]
- Biz/ed (Business Education on the Internet)[4]
- IHR-Info (Institute of Historical Research)[5]
- OMNI (Organising Medical Networked Information)[6]
- SOSIG (Social Science Information Gateway)[7]

The system is being used or evaluated by a number of other projects in the UK and interest in its use has also been expressed outside the UK. In addition, ROADS (along with SOSIG) is involved in the EU funded DESIRE[8] project (under the Fourth Framework Telematics Programme).

Each of the eLib ANR gateways is building a subject specific catalogue of Internet resource descriptions. The ROADS software attempts to be as modular as possible to allow the gateways to configure the 'look and feel' of the services; for example, in the way they present browsable listings of resources and search results. It also allows the gateways to pick and choose parts of the system appropriate to their service and 'plug in' their own applications; SOSIG plans to use this capability to add a thesaurus tool to the standard ROADS search facility.

Each service may also differ on issues of selection policy, classification, etc. However, the gateways share a common metadata format for collecting information and end-users

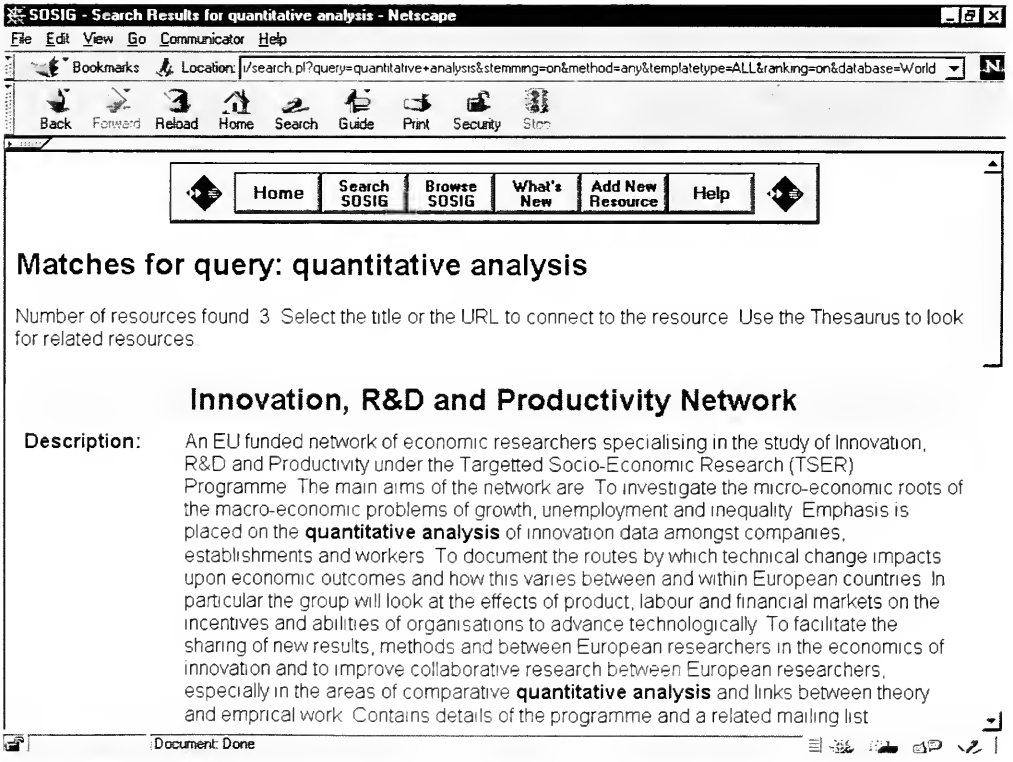
will ultimately be able to cross-search the gateways (using the WHOIS++ directory technology).

ROADS Templates

In addition to software development, the ROADS project is concerned with issues of metadata. The eLib ANR gateways are creating records for selected quality resources on the Internet using a standard template. The metadata format used by ROADS is based on the Internet Anonymous FTP Archive (IAFA) Template definitions[9]; which as the name suggests were originally designed to describe resources available through FTP sites. With the growth of the Web, these templates were extended to cover Internet resources in general. The templates have been extended further by ROADS based on the implementation experiences of the subject gateways (therefore the templates will be referred to throughout the paper as ROADS templates rather than IAFA). The template format was originally designed to be created by site administrators and therefore the emphasis is on simplicity and ease of creation. This simplicity also means that information skills are not essential and subject gateways can use the expertise of subject specialists as well library and information professionals to build their catalogues.

The template is a text-based record composed of a series of attribute-value pairs that describe a resource content, format and location. A number of different resource description template types exist:

- DOCUMENT
- IMAGE
- MAILARCHIVE
- ORGANIZATION
- PROJECT
- SERVICE
- SOFTWARE
- SOUND
- USENET
- USER



Example of search results in SOSIG

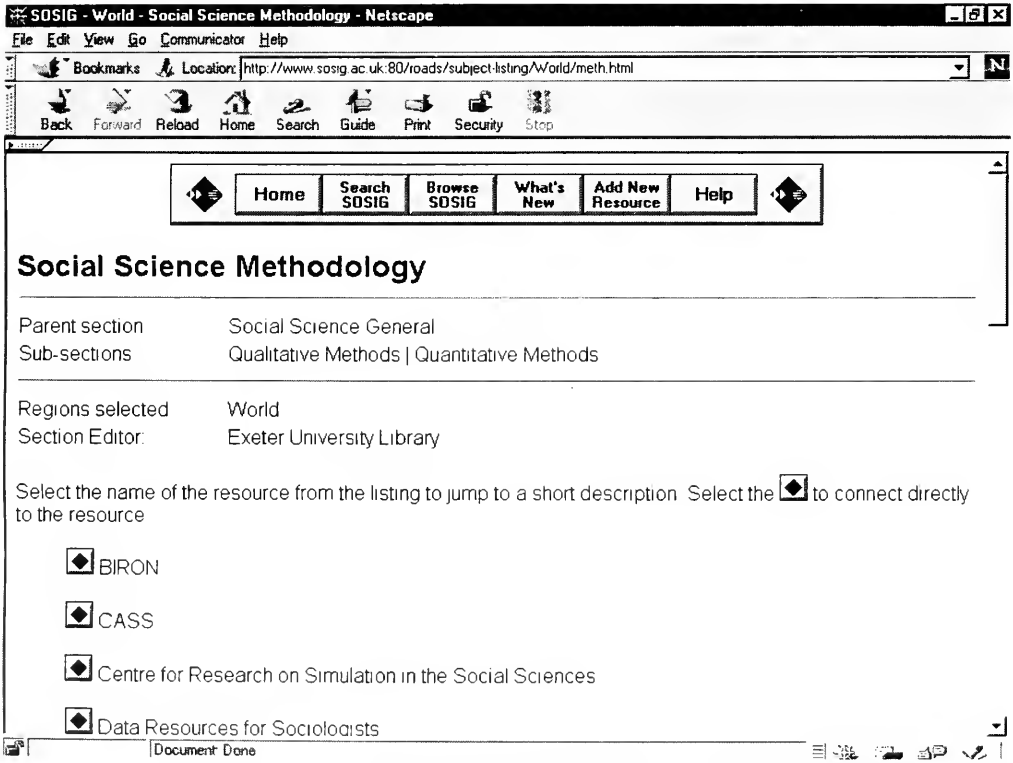
■ VIDEO

Within a ROADS template, there are three kinds of attribute; plain, variant and cluster. Plain attributes describe the basic characteristics of a resource, such as Title, Description and Keywords. They contain information about a resource that is only required once. Variant attributes are repeated for multiple versions of a resource. Examples of variant attributes are Language and URI; if a document is available in English and French two sets of variant attributes would be used to record the language and URL of each version. Other examples of variant attributes include format and size of the resource. Cluster attributes record information that may be common to a number of resources, for example name, address and email details of individuals or organisations.

The original IAFA templates have been extended slightly based on requirements from the subject gateways. For example, the attributes Subject-descriptor and Subject-descriptor-scheme have been added. These attributes allow the resources to be classified using an appropriate classification scheme and this information is used to form the basis of browsable listings on the gateways. A range of administrative attributes was also added.

The subject gateways can choose which template types and attributes to use according to the requirements of their end-users. However, a minimum set of attributes exists to ensure a level of interoperability between the gateways. The core attributes are: Title, Description, Keywords, URI, Subject-descriptor and Subject-descriptor-scheme. This does not include any of the administrative attributes such as the record creation date as these are generated automatically. Individual gateways may choose to make other attributes mandatory according to the requirements of their user community. For example, SOSIG is currently extending its coverage of European resources and has made Language and Country mandatory attributes in order to support this.

A registry of the templates[10] is maintained at UKOLN.



Example of browsable listing in SOSIG

This allows ROADS gateways to register the need for new template types or new attributes within existing templates if or when required. The registry will also contain some basic cataloguing rules to assist with interoperability.

Example of a ROADS Template

Template-Type: SERVICE
Handle: SOSIG472
Category: Database
Title: IBSS ONLINE
Alternative-Title: International Bibliography of the Social Sciences
URI-v1: telnet://bids.ac.uk
URI-v2: http://www.bids.ac.uk/ibss
Admin-Handle-v1:
Admin-Name-v1:
Admin-Work-Postal-v1:
Admin-Country-v1: uk
Admin-Work-Phone-v1: +44 (0)1225-826074
Admin-Work-Fax-v1:
Admin-Job-Title-v1:
Admin-Department-v1:
Admin-Email-v1: bidshelp@bids.ac.uk
Publisher-Handle-v1:
Publisher-Name-v1: Bath Information & Data Services
Publisher-Type-v1:
Publisher-Work-Postal-v1: University of Bath, Bath, BA2 7AY
Publisher-Country-v1: uk
Publisher-Work-Phone-v1:
Publisher-Work-Fax-v1:
Publisher-Email-v1:
Description: IBSS ONLINE provides electronic access to the database of the International Bibliography of the Social Sciences. It contains the bibliographic details of journal articles, book reviews books, and the chapters from selected multi-authored monographs. The database contains over 680,000 records covering publications appearing between 1981 and the present day, and is growing at the rate of approximately 100,000 items per annum. Subject coverage is based on the four principal disciplines of anthropology, economics, political science and sociology, but it also reflects the interdisciplinary nature of the social sciences. Material can be found which covers, for example, agriculture, archaeology, business studies, criminology, education, environmental issues, history, law, social policy, social work, and statistical methods. There is extensive coverage of international material. Records come from over 100 countries, and 95 different languages are represented in the database. The database is mounted at Bath Information & Data Services (BIDS). All members of UK higher education institutions (HEIs) funded by the HEFCs are eligible to use IBSS ONLINE free at the point of use. Users must register with their own institution's library.
Keywords: social science, sociology, politics, economics, anthropology
Authentication: Access to most databases is by username and password.
Registration: Users are required to register with a representative at their own HE institution.
Access-Policy: The IBSS ONLINE data may only be used by an employee, student or other person authorised by the institution or organisation which has taken out a licence to use the service.
Access-Times: All BIDS services are normally available 24 hours a day, 7 days a week.
Copyright:
Subject-Descriptor-v1: 3,301,32,33,572
Subject-Descriptor-Scheme-v1: UDC
Language-v1: en
Language-v2: en
ISSN:
Source:
To-Be-Reviewed-Date:
Record-Last-Verified-Email:
Record-Last-Verified-Date:
Destination: UK,WORLD
Record-Last-Modified-Date: Mon, 28 Apr 1997 17:21:46 +0000
Record-Last-Modified-Email: ecdh@aubergine.ilrt.bris.ac.uk
Record-Created-Date: Wed, 15 Jun 1995 13:22:00 +0000
Record-Created-Email: ecdh@ssa.bris.ac.uk

Mapping ROADS Templates

One of the main project objectives of ROADS is to 'implement and test emerging standards and to improve UK participation in international standards making activity'[11]. To this end the project closely monitors metadata developments and is very active in metadata standards initiatives, in particular the Dublin Core and Warwick Framework. Mapping ROADS templates to WHOIS++ templates has been done as part of the planned developments for distributed searching of ROADS databases (these are actually very similar in format to the ROADS templates). In addition UKOLN has produced several textual mappings from the ROADS templates to other metadata formats such as USMARC, Dublin Core, SOIF and the Z39.50 Bib-1 attribute set[12]. The templates map reasonably well on to these other formats although there may be some difficulties with syntax.

It would seem fair to assume that there will continue to be several metadata formats in use to describe Internet resources. However, ROADS has committed to providing

conversion tools if they are required by the eLib gateways and some proof of concept work has already been carried out converting the ROADS templates into USMARC and other formats. As part of another project, an experimental Z39.50/WHOIS++ gateway has been built which allows users to search a ROADS database in parallel with a range of Z39.50 databases[13].

ROADS Developments

ROADS is presently in version 1 of its development cycle; this provides all the tools and software to build and maintain a gateway of Internet resources. ROADS version 2 (already in alpha development) will continue to enhance these tools in response to requirements from the subject gateways. In addition to this, the next version will include:

Cross Searching Distributed Databases

Currently each ROADS subject gateway is searchable in a standalone format although some experimental work on cross searching has already been carried out between some of the gateways. The project is using the WHOIS++ search and retrieval protocol developed by Bunyip Information Systems (who provided some industrial consultancy on the project). This allows distributed databases to be queried over the network and the next version of the software will fully support this cross searching mechanism. In addition to linking the databases together, the project will be investigating the use of a related technology - the Common Indexing Protocol (CIP) to provide a method of routing search queries to appropriate databases[14].

Cross searching will be particularly useful for the SOSIG and Biz/ed gateways whose subject areas (social sciences and business and economics) overlap; potentially causing confusion for end users trying to identify which gateway they should use. Once cross searching is implemented, SOSIG will no longer continue to catalogue business or economics resources but users of the SOSIG gateway will still be able to search for and find economics related resources. Because of the overlap, the two projects are also looking at ways of presenting browsable lists across the two projects.

As part of the DESIRE project SOSIG is also hoping to collaborate with social science institutions or libraries in Europe who want to set up national databases of networked resources. European institutions would be able to make use of the tools and documentation developed by ROADS and DESIRE to create national gateways. This model is currently being piloted by the Koninklijke Bibliotheek (National Library of the Netherlands) who are building a ROADS database of Dutch social science resources.

Harvesting Resources

The eLib ANR gateways concentrate on cataloguing high quality Internet resources and it is this human input that distinguishes them from other Web search tools such as AltaVista. Users of the gateways are not overwhelmed by thousands of matches to their queries but are presented with a small number of resources which have been through a careful process of selection and description. However, this process means that there is a high cost associated with the creation of the catalogue records. Consequently, the gateways tend to catalogue at a server level rather than at the level of individual documents or pages. ROADS is looking at incorporating a Harvest-type technology in order to try to bridge the gap between the 'hand picked' approach of the gateways and the so called 'vacuum cleaner' approach of the Web search engines. One approach is to use a harvested database to supplement the quality-catalogued records and ROADS is investigating ways to integrate and present the two.

One of the aims of ROADS and of the subject based gateways is to 'encourage information providers to become involved in the creation of records about their own data in order to make their information as useful and accessible as possible'[15]. Typically, information providers supply little or no metadata with their resources, due in part to a lack of standards or direction in this area. ROADS is promoting the idea of 'Trusted Information Providers' (TIPS) who would be identified by the individual subject gateways. The TIPS may be services or institutions whose information had been previously validated by the gateways that would provide metadata with their resources to be collected automatically.

This second level of approach to support the TIPS idea is to develop a tool that can be used to pre-populate ROADS databases by harvesting metadata from resources and inserting them into templates. The cataloguers can then 'add value' to the automatically generated template before finally submitting it to the database. The ROADS Harvester can be used to generate a single template based on one URL or it can be run recursively across a range of URLs as a 'Bulk Harvest'. The Harvester is still under development but in the longer term should help the gateways to redress the imbalance between quantity and quality.

Contact Details

Debra Hiom is a Research Officer on the SOSIG and DESIRE projects at the Institute for Learning and Research Technology, University of Bristol in the UK. She can be contacted at the following address:

Institute for Learning and Research Technology,
University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK.
Tel: +44 (0)117 928 8443
Fax: +44 (0)117 928 8478
Email: D.Hiom@bristol.ac.uk

Acknowledgements

This paper references the work of the ROADS project team; in particular Jon Knight and Martin Hamilton at Loughborough University, Rachel Heery, Michael Day and Andy Powell at UKOLN and Chris Osborne and Paul Hofman at the University of Bristol. Any inaccuracies are the author's own.

For More Information About ROADS
If you would like more information about the ROADS project and availability of the software, contact Paul Hofman at: <roads-liaison@bris.ac.uk>

References

1. Resource Organisation and Discovery in Subject-based services

<URL: http://www.ukoln.ac.uk/roads/>

2. Electronic Libraries Programme

<URL: http://www.ukoln.ac.uk/elib/>

3. ADAM (Art, Design, Architecture and Media Information Gateway)

<URL: http://www.adam.ac.uk/>

4. Biz/ed (Business Education on the Internet)

<URL: http://www.bized.ac.uk/>

5. IHR-Info (Institute of Historical Research)

<URL: http://ihr.sas.ac.uk/>

6. OMNI (Organising Medical Networked Information)

<URL:http://www.omni.ac.uk/>

7. SOSIG (Social Science Information Gateway)

<URL:http://www.sosig.ac.uk/>

8. DESIRE

<URL:http://www.nic.surfnet.nl/surfnet/projects/desire/>

9. Publishing Information on the Internet with Anonymous FTP

<URL:http://www.roads.lut.ac.uk/System-docs/Internet-drafts/draft-ietf-iiir-publishing-03.txt>

10. ROADS Template Registry

<URL:http://www.ukoln.ac.uk/roads/templates/>

11. ROADS Objectives

<URL:http://www.ukoln.ac.uk/roads/

12. Mapping between metadata formats

<URL: http://www.ukoln.ac.uk/metadata/interoperability/>

13. Zexi: Z39.50 Experimental Implementation

<URL: http://www.roads.lut.ac.uk/zexi/>

14. The Common Indexing Protocol

<URL:http://www.roads.lut.ac.uk/System-docs/Internet-drafts/draft-ietf-find-cip-new-00.txt>

15. ROADS Objectives

<URL:http://www.ukoln.ac.uk/roads/>

* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

The Statistical Metadata Repository: an electronic catalog of survey descriptions at the U.S. census bureau

*by Daniel W. Gillman and Martin V. Appel**

1. Introduction

The U.S. Census Bureau (BOC) is developing a prototype statistical metadata repository for use with Internet data dissemination and automated integrated survey processing tools. The repository will be an electronic catalog of information about survey designs, processing, analyses, and data sets. Access will be through the Internet and the World Wide Web. Substantial background work was done before work to build the prototype could begin.

Statistical metadata is the information and documentation needed to describe and use statistical data sets for the lifetime of the data. The efficient, effective, electronic management of metadata greatly increases the usefulness of those data sets, especially for Internet data dissemination. Statistical metadata can also be used to facilitate survey design, processing, management, and analysis. Automated integrated survey processing systems, which create and use this information, will allow statistical agencies to conduct their programs in ways that were not possible before.

The repository is being designed based on standards and data models. It is being implemented as a relational database and organized through these standards and models. International, American, and internal Census Bureau standards are all being brought to bear in the development of the repository. Three models have been developed and integrated to form the structure of the repository. The models are the Business Data Model, the Data Element Registry Model, and a Metamodel.

Tools for the collection of the metadata and querying the repository are under development. Without the cooperation of the survey designers and analysts who create the metadata, the repository will never be populated. General, intuitive, and easy to use tools must be developed to collect the data. Conversely, the information in the repository will not be useful if it cannot be retrieved in an easy way. A survey Business Process Model, or table of contents, has been developed for users and analysts to find the type of information they may want to provide. This table of contents is being used as a template in the design of the tools. Also, it can be used to design a low level interface for other systems to access and use the repository.

Substantial benefits should be available to the Census Bureau when the repository is functional. It organizes the documents, data sets, and variable descriptions of the agency. The repository will allow for comparisons across surveys (data or designs) which previously have not been easily available. Finally, the repository will make the public information of the agency fully available from a common source. If other statistical agencies around the world adopt similar approaches, the concept of a "single world-wide statistical agency" on the Internet could become reality.

This paper will define what statistical metadata is, describe the design of the repository (including the standards and models), describe the tools under development for populating and querying the repository (including the table of contents outline), and discuss the ramifications for the agency of implementing the repository.

2. Definitions

Statistical Metadata is descriptive information or documentation about statistical data, i.e. microdata and macrodata. Statistical metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data.

The two types of statistical data (electronic or otherwise) are described as follows (see Lenz, 1994):

- Microdata - data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment.

- Macrodata - data derived from microdata by statistics on groups or aggregates, such as counts, means, or frequencies.

The extensive nature of statistical metadata lends itself to categorization (see Sumpter, 1994) into three components or levels:

- Systems - the information about the physical characteristics of the application's data set(s), such as

location, record layout, database schemas, media, size, etc;

- Applications - the information about the application's products and procedures, such as sample designs, questionnaires, software, variable definitions, edit specifications, etc;

- Administrative - the management information, such as budgets, costs, schedules, etc.

The systems, applications, and administrative components help to differentiate the sources and uses of statistical metadata. Some authors (see, for example, Sundgren, 1991b, 1992, 1993) refer to the applications and administrative components of metadata as meta-information. We chose to use the term metadata because it seems to simplify the discussion.

Statistical metadata and metadata repositories have two basic purposes (see Sundgren, 1991a, 1991b, 1992, 1993):

- End-user oriented purpose: to support potential users of statistical information, e.g. through Internet data dissemination systems; and

- Production oriented purpose: to support the planning, design, operation, processing, and evaluation of statistical surveys, e.g. through automated integrated processing systems.

A potential end-user of statistical information needs to

- identify,
- locate,
- retrieve,
- process,
- v interpret, and
- analyze

statistical data that may be relevant for a task that the user has at hand.

The production-oriented user's tasks belong to the following types of activities:

- planning/design/maintenance,
- implementation/processing/operation, and
- v evaluation.

An input-oriented statistical agency is one where the statistical surveys they conduct or manage are also the natural building blocks of its organization. The BOC is currently an example of such a statistical office.

An output-oriented statistical agency is one which focuses on meeting the needs of its customers. The BOC is striving to become more output-oriented. See Sundgren (1991a, 1991b, 1992, 1993) for a more detailed discussion of these ideas. Output-oriented database systems relate data from different surveys. They need special software and metadata tools for reconciling data from different sources and for helping the users to interpret and analyze the data. This paper describes the pieces necessary to build those metadata tools.

Statistical Metadata Repository (MDR) is a planned repository of statistical metadata and pointers to other metadata (such as documents or images). A proof-of-concept system has been built (see Gillman and Appel, 1994), and a series of prototypes are under development. The design, uses, and functionality of the MDR will be discussed in more detail below.

3. Statistical Metadata Repository

The MDR is being designed to assist with two new types of tools which are under development at the BOC: Internet data dissemination ; and automated integrated survey processing systems . These tools correspond to the end-user oriented purpose and production oriented purpose, respectively, of statistical systems. Statistical systems are known formally as Statistical Information Systems (SIS) (see Sundgren, 1991b, 1992, 1993; or Gillman, Appel, and LaPlant, 1996).

3.1 Purposes

The eventual plan for the MDR is that it will contain the metadata for survey designs, processing, analyses, datasets, and related information for all surveys the BOC performs. Links to the data files, documentation, and images (such as questionnaire forms) will also be stored (see Sundgren, et al, 1996; or Appel, et al, 1996).

This has led to the management of data in a decentralized and non-uniform way. On one hand, there is a need for the survey management to process and manage their data in the most efficient way. On the other, there is a need for data users to be able to find and access data efficiently and effectively. The MDR will facilitate a solution for the data users while allowing the survey data managers to find a smooth transition to standard data management strategies.

There are many functions for which the MDR is being designed. Primarily, the MDR will be a standard tool for researchers and analysts to locate survey data and metadata. Data dictionaries, record layouts, questionnaires, sample designs, and standard errors are examples of information

that will be directly available. Links from subject types, e.g., income, race, age, and geography, to data sets will allow users to locate data sets by subject. Less obviously, users can compare designs of different surveys and find common information collected by them.

The MDR will help facilitate data administration at the BOC. Many surveys define data elements with the same name but with (slightly) different definitions. An aim of the MDR is to help people manage this problem. If definitions and other attributes of data elements are standardized across surveys, through the use of a data element registry (a subset of MDR), then confusion generated by the differences in meaning will be reduced. Naming standards and conventions are also needed to reduce the confusion. The MDR will provide the information necessary for the user to understand the distinctions and similarities among data elements from multiple data sources. The design of the data element registry part of the MDR will be based on a standard, and it will be discussed in more detail in section 3.2.2.

Many of the purposes for the MDR are associated with both the end-user orientation and production orientation. Here we will list the end-user oriented purposes. The typical end-user oriented SIS is an Internet data dissemination system. Some of the major functionality for the MDR in support of this is:

- Location of data sets by survey name and date or content (e.g. household income);
- Names, definitions, and related information about data elements and links to the surveys and data sets that use them;
- Links to documentation describing aspects of survey design, processing, or analysis;
- Links across documents to identify common themes contained in them;
- Links to images (e.g. questionnaire forms) that are of interest;
- The ability to search the information potential through query languages such as SQL.

The typical production oriented SIS is an automated integrated survey processing system. Most of the purposes of the MDR for the end-user oriented systems are common to the production oriented systems as well. Often, production oriented SIS users will be survey analysts working within the BOC (statistical agency). They have and need access to confidential data to which external end-users cannot have access. The additional functionality must support this use, such as:

- Links to all the data sets produced by the instance of a survey (e.g. Current Population Survey, June 1996);
- Links to frame, sample, and administrative records files;
- Links to a management information system;
- Links to some confidential metadata such as disclosure analysis algorithms.

These lists are not meant to be inclusive, but to give a fairly extensive picture of the potential uses for the MDR.

3.2 Models

The design of the MDR is based on three data models. Within the repository, these models have been integrated into one extensive model which covers many aspects of statistical metadata. Extensions to the model are planned as new items or needs are identified.

The three models represent the major dimensions to the MDR model (see figure 1). They are described briefly here and will be discussed in more detail below:

- Business Data Model - The model describes the business of the BOC - surveys. It describes survey designs, processing, analyses, datasets, products, and documents as related to statistical surveys.
- Data Element Registry Model - A data element registry is a mechanism for managing the names, definitions, permissible values, and other attributes of data elements. Metadata describing data elements is entered into the registry by a process called registration. Expanding the concept of registration to include surveys, products, datasets, and documents, this model handles the needs of registering metadata.
- Metamodel - This model describes application specific areas and other non-business related items such as security, access control, database schemas, record layouts, and time frames. The metamodel provides the repository's view to itself.

The MDR prototype also uses a business process model described below.

- Table of Contents - A business process model has also been developed. It is in the form of an outline, or table of contents (TOC). The TOC describes the processes of a survey from design to data dissemination.

The MDR model can also be divided into five functional areas. This view gives a clearer picture of how the integrated model works (see figure 2).

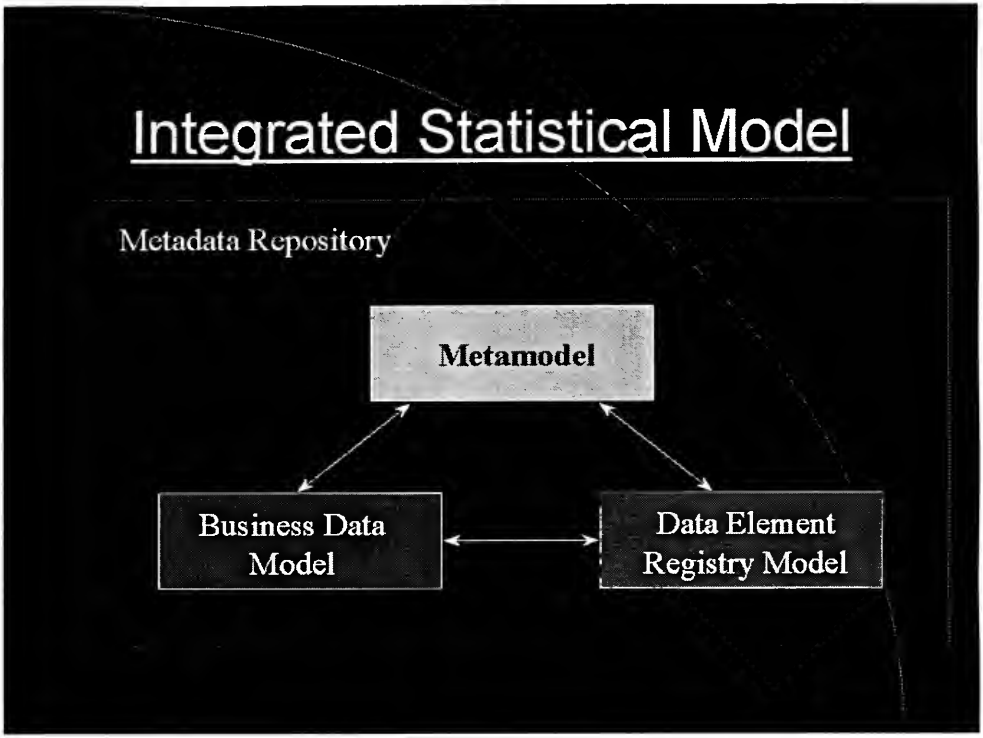


Figure 1: Overview of Integrated Model

The functional areas are:

- Data Element Registry - Manages the names, definitions, permissible values, and other attributes of data elements (see above and below).

- Registration - Manages the metadata needed to register items for which the repository keeps track: surveys, data elements, documents, datasets, products. This section handles the information types which are common to each of the objects which are registered in the MDR, much like an electronic card catalog system.

- Metamodel - Manages the application specific information such as security and access control, search criteria, record layouts, database schemas and access, etc (see above and below).

- Business Data - Manages information about surveys, including design, processing, and data (see above and below).

- Documentation - Manages information about documents. The association of documents to different records within other parts of the model acts as a classification system for the documents.

3.2.1 Business Data Model

The Business Data Model (BDM) describes the business (statistical surveys) of the BOC. It is composed of entities, attributes, and relationships which describe information that a statistical agency needs to keep about surveys. Much of this information is in the form of specifications or procedural documentation. The model supports the storage of metadata as single attributes or as documents. Figure 3 is a high level ER diagram of the BDM, and see Appendix

A for an entity definition list.

The BDM describes survey designs, processing, analyses, and datasets. It contains entities for each of the important parts of a survey: universe, frame, sample, questionnaire, etc. The model allows for the organized storage and search for metadata about a survey, and it allows searching for metadata items across surveys. Many statistical metadata systems in use today address the metadata needs for a single survey or application, but the BDM addresses the metadata needs for many surveys.

An important feature of the BDM is that documentation is handled in a general way. Each entity of the model allows for many documents to be attached to a single record. The documents can be distinguished by version, document type (e.g. specification, procedure, memo, etc.), the entity the document is associated with, and the relationships the given record has with other records in the model. This provides a comprehensive classification scheme for documents which helps users search directly for the information they need. Coupled with the indexed and key word search provided by most Internet search engines, the BDM is a powerful document management paradigm.

The model also provides several other features listed below:

- maintains a list of all current surveys conducted by the agency;
- allows for comparing designs, specifications, or procedures across surveys;
- allows for reuse of designs, specifications, or procedures;
- provides for assembling complete documentation for a survey.

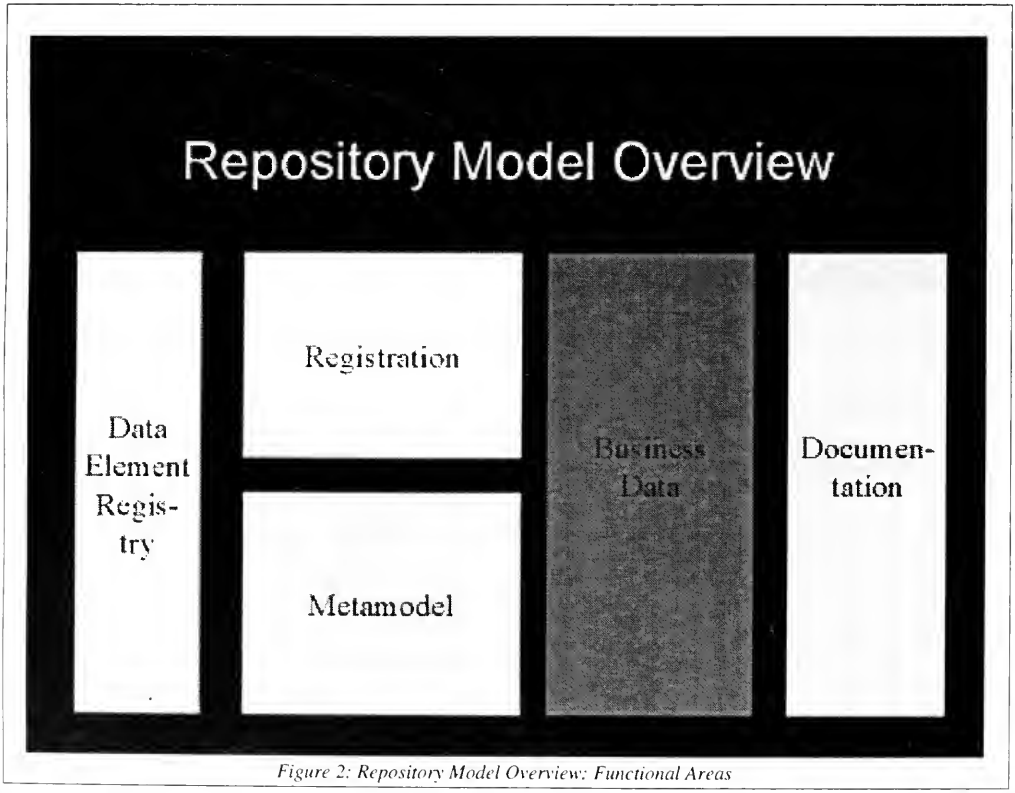


Figure 2: Repository Model Overview: Functional Areas

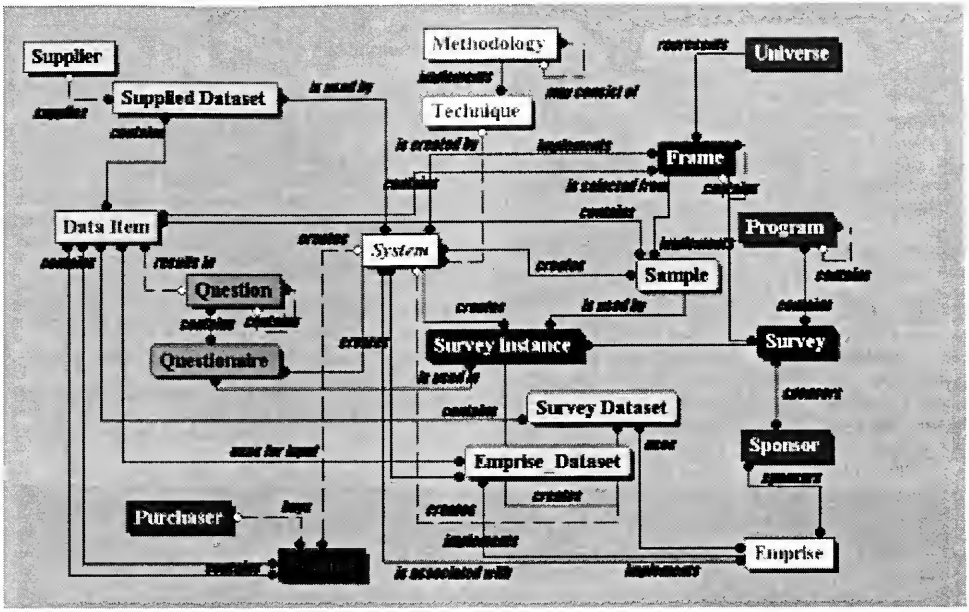


Figure 3: Business Data Model

3.2.2 Data Element Registry Model

Data elements (or variables) are the fundamental units of data an organization collects, processes, and disseminates. A data element registry (DER) is a mechanism for managing data elements in a logical fashion. DER's organize information about data elements, provide access to the information, facilitate standardization, help identify duplicates, and facilitate data sharing. DER's are like data dictionaries in that they contain definitions of data elements. But more than data dictionaries, they contain all the information about individual data elements that an organization requires. Data dictionaries are usually associated with single data sets (files or databases), but a DER contains information about the data elements for an entire program or organization. The information contained in a DER is part of an organization's metadata. Therefore, the registry itself will be part of the MDR.

Important applications for DER's include SIS's. Electronic data dissemination requires easy access to information about data elements. Data element names, definitions, and classification schemes will help users in locating and understanding data sets. Automated integrated survey processing systems that will include sample and questionnaire design, automated edits and imputation, and coding systems require full descriptions of data elements. Designers need to know the definitions of all variables that may be affected by the programs they are using.

The DER model provides for all the metadata needed to describe data elements. It also provides the entities necessary for registration and standardization of data elements. Generalizing the concept of registration (see section 4.2 below) to include documents, datasets, products, and surveys provides a framework for merging the DER and the BDM. A consequence of registering the important metadata items in the MDR is that the repository, from the registration point of view, becomes a card catalog of metadata items. The integration must also include linking data elements to each of the entities in the BDM which use them (e.g. frame, sample, survey dataset, question, etc.).

An important feature of the DER is that data elements are composed of a concept (data element concept) and a representation or value domain (set of permissible values). The power of this is seen as follows:

- sets of similar data elements are linked to a shared concept, reducing search time;
- every representation associated with a concept (i.e.

each data element) can be shown together, increasing flexibility;

- all data elements that are represented by a single (reusable) value domain (e.g. SIC codes) can be located, assisting administration of a registry;

- similar data elements are located through similar concepts, again assisting searches and administration of a registry.

See figure 4 for a high level ER diagram of the DER, and see Appendix B for an entity definition list.

3.2.3 Metamodel

The metamodel is the repository's view of itself. It contains application specific entities necessary for the functioning of particular SIS's, and information which controls access to metadata in the rest of the repository. The kinds of information the metamodel handles are access control, security, physical location of data, machine addresses, record layouts, database schemas, access procedures, etc.

The development of the metamodel has been iterative. No specific metamodel has been built. Instead, as new functions are identified, they have been added to the MDR model. The partnerships (see section 3.4) that have been formed with SIS developers for using the MDR model have been a rich source for metamodel entities and attributes. As these partnerships continue and the SIS's are further developed, more information is added to the metamodel and to the MDR model.

.3.2.4 Business Process Model

A table of contents (TOC) outline view (see Census Bureau, 1996) of survey processes has been developed. It was patterned after work done by a BOC Reinvention Lab and at Statistics Sweden (see Rosen and Sundgren, 1991). The TOC is formally a Business Process Model. It is

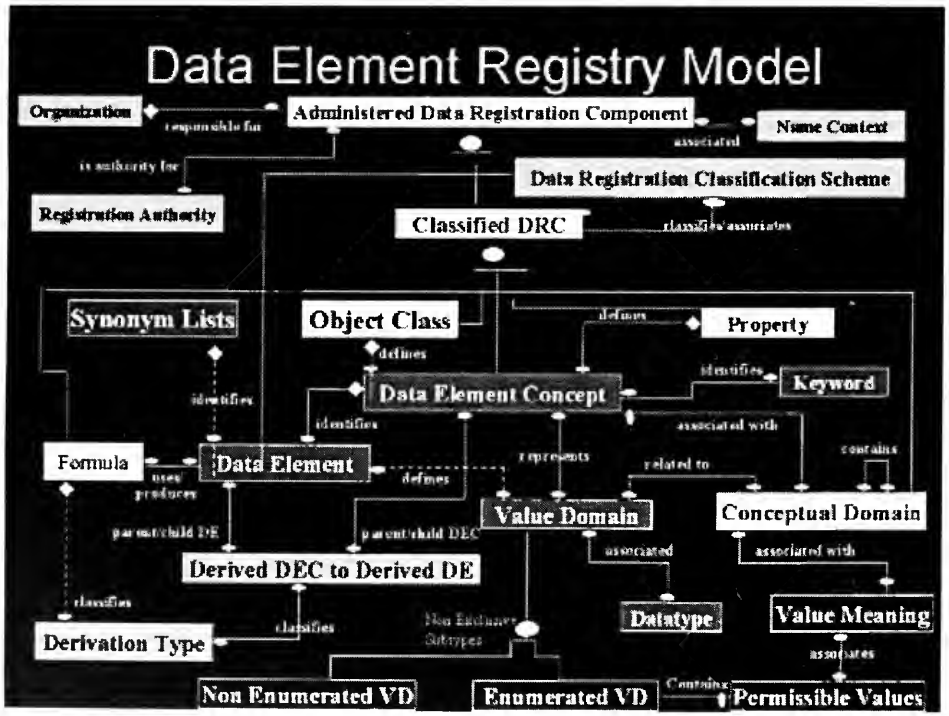


Figure 4: Data Element Registry Model

divided into eight chapters, each detailing a different aspect of survey processing. The chapter names and their descriptions follow below:

- Content - The Content refers to the nature of the information that is the subject of the survey, i.e. what the universe is, a description of the data collected, and a description of the resulting products. May contain definitions, and data standardization and coding information.

- Planning - Documentation related to the planning and management of the design; the conduct of the survey and the analysis, dissemination and disposition of the data. This includes documentation related to budgeting, manpower, and training.

- Design - The design and specifications for how the survey will be conducted. Includes the design of the frame, sample, and questionnaire; and the specifications for edits, coverage, and estimations.

- Data Collection - Obtaining information from respondents and the conversion of that data into a form which can be processed.

- Data Processing - The stage of a project, following collection and receipt of the original material and preceding report-writing, during which the information is entered onto a machine-readable medium (or directly into a computer system) and eventually used to produce tabulations and statistical analyses.

- Data Analysis - Documentation related to all statistical processes used to analyze the survey results or those used for displaying or presenting the resultant information.

- Data Dissemination - The process of making data available to users, electronically or otherwise. Electronic data dissemination includes use of the Internet or CD-ROMs.

- Data - Any information gathered as the result of a survey or added to a survey form.

There are two uses that are being developed for the TOC: 1) to be used as a "check list" for users who need to provide metadata or users who want to search metadata from the MDR (see section 4.2); and 2) to serve as a mapping between the MDR and other repositories which need to share metadata (see Gillman, Appel, and LaPlant, 1996). In particular, the TOC can be used as a means to classify documents from another repository in the MDR.

3.3 Standards

In this section the applicable standards which have been used to guide the development of the MDR and its associated tools will be described briefly.

3.3.1 Data Element Standards

The model for the data element registry portion of MDR is based on the conceptual framework contained in the ANSI draft standard, The Metamodel for the Management of Shareable Data (MMSD), ANSI X3.285. It, in turn, incorporates all the principles described in an emerging international standard, Specification and Standardization of Data Elements, ISO/IEC 11179 (see ANSI X3L8, 1996). ANSI X3.285 provides a conceptual model for building a data element registry and contains some extensions to the framework described in ISO/IEC 11179. A complete data dictionary describing all the entities, attributes, and relationships of the conceptual metamodel is included in this document.

The MMSD metamodel provides a detailed description of the types of information which should belong to a data element registry. It provides a framework for how data elements are formed and the relationships among the parts. Implementing this scheme will provide users the information they need to understand an organization's data elements.

ISO/IEC 11179 is being developed in six parts. The names of the parts, a short description of each, and the status follow below:

- Part 1 - Framework for the Specification and Standardization of Data Elements - Provides an overview of the concepts in the rest of the standard. The current status of this document is Committee Draft.

- Part 2 - Classification of Data Elements - Describes how to classify data elements. The current status of this document is Working Draft.

- Part 3 - Basic Attributes of Data Elements - Defines the basic set of metadata for describing a data elements. This document is an International Standard.

- Part 4 - Rules and Guidelines for the Formulation of Data Definitions - Specifies rules and guidelines for building definitions of data elements. This document is an International Standard.

- Part 5 - Naming and Identification Principles for Data Elements - Specifies rules and guidelines for naming and designing non-intelligent identifiers for data elements. This document is an International Standard.

- Part 6 - Registration of Data Elements - Describes the functions and rules that govern a data element registration authority. This document is an International Standard.

3.3.2 Survey Design and Statistical Methodology Metadata Content Standard

The Survey Design and Statistical Methodology Metadata Content Standard (SDSM) (see LaPlant, et al, 1996; or Census Bureau, 1997) is a draft statistical metadata content standard for the BOC. It will provide a description of the information or documentation about statistical data. The content and design of the standard is based primarily on the BDM. The entities of the BDM specify the content sections of the SDSM.

SDSM will provide developers and users of statistical products with a common vocabulary for describing the design processing, analysis, and data sets for censuses and surveys. The SDSM also will serve as a glossary of statistical metadata concepts. Broad agreement on the meaning and organization of these concepts will provide the basis for improved communication among the producers and users of economic and demographic statistical data sets.

Each of the 29 sections in the SDSM consists of a list of entries, some that reference other sections. Each entry is a metadata data element. Any of these metadata data elements may be used to identify specific instances of metadata. The metadata may be some specific information (such as a number or text) or a url to a file of some type (e.g. documents, gif's, etc.)

The SDSM has been submitted to the formal standards review process of the BOC, and is expected to be issued as a BOC standard in Summer 1997. Once this occurs, it is hoped that other statistical agencies will adopt the SDSM or similar standards.

3.3.3 Other Standards

Information Resource Dictionary System (IRDS) is a standard which addresses the use, control, organization, and documentation of the information resources of an enterprise (see NIST, 1989). It is an application of another standard, Reference Model for Data Management (RMDM) (see ISO, 1995). The organization of the MDR model is based on the organization specified in IRDS. See Graves and Gillman (1996) for a more detailed discussion.

The Federal Geographic Data Committee (FGDC) of the U.S. Government has developed a family of metadata standards which addresses the geographic content of data. Executive Order 12906 has mandated that all U.S. agencies that produce geographic based data use these FGDC standards. Most BOC data is based on geography, therefore these standards will apply to BOC data.

Government Information Locator Service (GILS) (FIPS-192) is an extensible standard which describes a format and specifies the underlying protocol (NISO Z39.50) for making metadata available on the Internet. Another Executive Order (the Paperwork Reduction Act of 1995) mandates that all U.S. agencies create and maintain GILS records. This provides a mechanism for the public to find information about what their government is doing and producing through electronic means.

3.4 Partnerships

Several groups within the BOC developing SIS's have agreed to use the MDR structure to support the underlying metadata needs of those systems. A short description of each SIS follows below.

DADS (Data Access and Dissemination System) is the name for the Census Bureau initiative to develop and implement data access and dissemination focused on the 2000 Decennial Census and Continuous Measurement data sets, but with the ability to accommodate other data sets having geographic detail, such as those produced from the Economic and Agricultural Censuses.

The main objective of DADS is to provide one general (electronic) system for all access to Census Bureau data. The system will be designed to be fast, flexible, and cost-efficient. To achieve this, four cross-directorate teams were formed to study and recommend policies or designs for user input, promotion and outreach, pricing for products, copyright or trademark, corporate look and feel, data archiving, metadata and documentation, and coordination of various efforts and activities. DADS will attempt to incorporate other work, such as FERRET, where it is appropriate.

The DADS team is following a schedule to produce a new prototype each year, in the month of September, until the full production system is built in 2002. The 1996 prototype was a success, though limited in scope. The MDR model will be used to organize the metadata for the 1997 prototype.

FERRET (Federal Electronic Research and Review Extraction Tool) (see Capps, 1995) is a data extraction tool available on the Internet that allows users to find information about monthly demographic survey data using a World Wide Web browser. Users can select microdata items (individual survey question items) which can be used to create custom data queries. In addition, users can select macrodata (aggregated or summarized) tables to get preformatted survey data. Results of data queries can be output in SAS datasets or ASCII files. These results can be viewed on the screen or can be downloaded to a local computer. The SAS output allows one to get the results in pie charts, bar charts, or summarized on a U.S. map. The ASCII output can be brought into an Excel spreadsheet.

The FERRET system can be divided into four major parts. The first part is the user interface which is via the World Wide Web. The Ferret repository contains metadata such

as basic variable definitions, keywords, concepts, and other items. The Document Management System handles the documents which describe the survey design, processing, and analysis. Finally, there are two databases handling all the microdata and macrodata.

FERRET currently handles Current Population Survey data. Plans are to add other demographic survey data in the future. Work is also underway to make the FERRET repository model and the MDR model compatible. This will enable people to work with DADS and FERRET systems seamlessly.

StEPS (Standard Economic Processing System) (see StEPS, 1996) is an integrated survey processing system the objective of which is to eliminate redundant processing by combining existing survey systems into one system. The scope of the StEPS system includes providing the following basic survey processing functions:

- Data review and correction;
- Edits;
- Imputation;
- Outliers;
- Estimation;
- Estimation variance;
- Disclosure analysis;
- Time series;
- Queries (canned/ad hoc);
- Tables (canned/ad hoc);
- Management information; and
- Survey control operations (for scheduling of batch mode processes).

It will also provide the following additional functions:

- Generate standard and non-standard mail files for mail-out operations;
- Generate standard telephone files for telephone follow-up operations;
- Maintain standard variable names and flags;
- Maintain standard data structures;
- Allow entry of survey design specifications including edit and imputation parameters as determined by analysts or through automated historical data analysis
- Provide audit trails and backup capabilities;
- Provide access to SSEL; and
- Provide access to other economic area surveys and censuses.

The above provides a view of the functionality which StEPS will be designed to provide. Implementation details are not yet available. The StEPS system developers plan to use the MDR as a source of information about variables.

Product Registration is a multi-divisional effort to unify the systems that manage the production, inventory, distribution, and sale of Census Bureau products. The MDR model will be used to register products, i.e. link products to the variables, surveys, geography, and other items that will enable users to locate them. This work has recently started.

3. Metadata Management

The main aspects of managing metadata are content, storage, collection, registration, retrieval, system integration, and metadata administration. This section will describe how the standards based approach and the proposed design architecture address each of these aspects.

4.1 Content and Storage

Content refers to the identification of which metadata will be collected and stored in the MDR, and Storage refers to the how, i.e. the physical and logical mechanisms for storing the metadata. Much of the paper to this point has been addressing these issues.

The prototype MDR is being built using Oracle RDBMS as its underlying storage mechanism and is based on the models and standards discussed above. The models and standards describe the metadata content and how that content is organized for storage.

4.2 Collection

Metadata collection is recognized as a very difficult problem because of the fundamental changes that the survey design and analyst teams must go through to perform their work. At the BOC and other statistical agencies, metadata (mostly documents, often in the form of memos) is created either electronically or on paper for each survey, but it is just beginning to be stored in an organized repository, database, or document management system. Asking people to use a new system to capture this metadata and organize it represents a big change. The tools that are created must mimic as closely as possible the working paradigm already in place, such as the use of certain word processors and templates for creating documents. A major problem is that the working paradigm for each survey design and analysis team is different. So, creating common tools will require substantial planning. Also, incentives must be found so that the designer/analysts will want to provide the metadata to the MDR. No matter how well designed, tools without an obvious payoff to the user will not be used. Management can help with the adoption of metadata collection tools by supporting their use, but the end-users will ultimately decide their fate.

4.3 Registration

Registration is the process of providing the MDR with its knowledge about the metadata, e.g. name, location, type, etc. The general classes of items which need to be registered are data elements, surveys, products, datasets, and documents. Registration requires several things:

- all the necessary attributes are specified;

- all the necessary links are made (e.g. linking a dataset to all the data elements in its data dictionary);

- classifying the registered item.

Registration tools will have to be designed, probably one for each class of item. The tools will require a template for the user to supply the necessary attributes and make the links to other metadata as needed. Appropriate classification structures will need to be accessible through the tool so each item can be classified.

Useful classification schemes already exist which can be incorporated into registration tools, such as

- TOC;

- Themes as specified in the Cultural and Demographic Data Metadata draft standard of the Federal Geographic Data Committee;

- Thesauri from Statistics Canada and the University of Essex (U.K.).

It will be useful for the BOC to build a taxonomy of statistical terms to help with the classification problem. Of course, effective classification schemes also help with the search for metadata and for understanding the semantics of data or metadata

Several prototype metadata collection tools are in place at the BOC and other statistical agencies. SCBDOK (at Statistics Sweden), Document Management System (DMS - in use with FERRET at BOC), and the commercial document management system PCDOC (for 1997 Economic Censuses) are all designed or being designed under the framework outlined above.

4.4 Retrieval

Retrieval refers to querying metadata in the MDR. Querying will be part of the design of General Purpose Browsers and of SIS's which work with the MDR. User interfaces for metadata-driven systems will let users query the metadata to locate data or other survey information. Query languages such as SQL will allow the user to retrieve any metadata which is in the MDR. Other search mechanisms such as WAIS, key word, and hyper-text are available through the Internet. This is especially important for documentation databases.

The TOC view of the SDSM can be used as a check list for categories of metadata. For users wishing to find information about surveys, searching the TOC for the appropriate subject (e.g. questionnaire design) will be useful. Since the TOC Process Model is designed to be a complete description of survey design, processing, analysis, and data sets, then the TOC view will provide users access to all the metadata the BOC has about a survey.

A prototype metadata browser for the MDR has been built, and browsers are being built for the DADS and FERRET data dissemination systems.

4.4 System Integration

In addition to the tools for collecting and querying metadata, the integration of the MDR with other SIS's needs to be seamless. Two general possibilities for accomplishing this exist. First, the TOC can be used.

A mapping exists between the TOC and the MDR model, and maps can be built from the TOC to the other SIS's by mapping the TOC to their metadata models. Then, a map will exist from the MDR to each SIS, through the TOC. The MDR will act as a hub, a central communication link between the different SIS's in use at the BOC (see Gillman, Appel, and LaPlant, 1996).

Another solution, probably more effective, is for developers of SIS's to adopt the MDR model for the metadata portion of the SIS. If every SIS at the BOC uses the MDR model, then a distributed metadata repository (each piece based on the same model) will be built. Tools designed to search the metadata in one SIS will be able to search the metadata in all SIS's. A seamless view of the metadata for the entire agency will result. Users who look for BOC data in FERRET will be able to locate data that is only accessible through DADS without having to know which tool to go to first. The actual viewing or downloading of the data will probably require switching tools, but that problem should be minor.

4.5 Metadata Administration

The adoption of the MDR model for storing metadata will require more than supplying information about data elements, surveys, documents, or datasets. Metadata administration is the active management of the information about all the agency's metadata. No function of this type exists at the BOC at this time at the agency level.

The registration process described in section 4.3, and the DER described in section 3.3.2, define generally the information that is required for accurate and complete data administration. The MDR model has expanded the notion of data registration to include metadata.

The registration tools discussed above will handle the entering of metadata into the MDR, but there is a human side to metadata administration which must not be lost in the discussion of the MDR. Some of these functions are:

- Determining which data elements have the same meanings as others;

- Determining whether metadata items have been properly classified;

- Ensuring all necessary information is properly supplied for each registered metadata item;

- Working with metadata administrators of other agencies to facilitate the sharing of data and metadata;

- Designing rules for forming metadata definitions.

- Designing and implementing naming conventions;

Metadata administration will require a large commitment from the BOC, but it will greatly enhance the usefulness of BOC data, make the MDR a better tool, and facilitate the sharing and understanding of data and metadata among groups within the BOC or with other agencies.

4. Prototype

A series of prototypes is currently under development. The first version is complete. It implemented a subset of the MDR model and contained some information about some data elements and documents. A browser tool was developed using the TOC as a search mechanism for specific types of documents. The browser is a Web based tool that uses a combination of basic HTML, CGI-Perl scripts, and JAVA.

The second prototype is under development now. It is expected to be complete in July. It will implement the complete MDR model, contain substantially more documents, and use an improved version of the browser. Two important functions will be demonstrated: the ability to find metadata across surveys and a tool to register metadata for products. Subsequent prototypes will add more functionality each time.

Usability testing is planned for some of the prototypes. Both the registration tools and the browser will require user feedback to ensure that the tools are useful for users. Unfortunately at this time, the prototypes cannot be released on the Web to the Internet. Much of the metadata in the MDR is not available for the public, and the security functions for the MDR have not been developed to the point where this information is secure.

5. Conclusion

This paper has discussed the work at the BOC to design and build a prototype statistical metadata repository (MDR) using standards developed by international, national, and U. S. Government organizations. Detailed data and metadata models have been built and integrated. The integrated model is the basis for the MDR architecture. It provides a structure for storing the metadata which describes survey designs, processing, analyses, and datasets. The model supports the card catalog metaphor for organizing the BOC metadata.

The MDR will not be an end in itself. Instead, it will work in conjunction with Internet data dissemination and automated integrated survey processing tools. Several examples of both of these tools are under development at the BOC. The MDR prototypes must be ready in time to meet the schedules of these other tools.

The first MDR prototype has been built and subsequent ones are planned. Registration and query tools are being developed, and the prototype MDR is being populated with metadata. Increasing interest in using the MDR model for storing metadata for various projects has increased the chance that a seamless distributed metadata repository for the BOC can be developed. Further research, planning, and work will be necessary to bring this plan to reality.

6. References

Appel, M. V., Gillman, D. W., LaPlant, W. P. Jr., Creecy, R. H. (1996), "Towards Unified Metadata Systems and Practices", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.

ANSI X3L8 - Data Representations (1996), "ISO/IEC 11179 Part 1 - Framework for the Specification and Standardization of Data Elements, Working Draft 7", February 1996.

Capps, C. (1995), "Overview of the Technical Architecture for FERRET", Census Bureau internal document, Demographic Surveys Division.

Census Bureau (1997), "Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, April, 1997.

Census Bureau (1996), "Table of Contents for Statistical Design and Survey Methodology Metadata Content Standard", Draft, Census Bureau Internal Document, July 2, 1996.

Gillman, D. W. and Appel, M. V. (1994), "Metadata Database Development at the Census Bureau", Presented at the UN/ECE METIS Working Group Meeting, Geneva Switzerland, November 22-25, 1994.

Gillman, D. W., Appel, M. V., and LaPlant, W. P. Jr. (1996), "Design Principles for a Unified Statistical Data/ Metadata System", Proceedings of SSDBM-8, Stockholm, Sweden, June 18-20, 1996.

Graves, R. B. and Gillman, D. W. (1996), "Standards for Management of Statistical Metadata: A Framework for Collaboration", ISIS-96, Bratislava, Slovakia, May 21-24, 1996.

ISO (1995), "Reference Model for Data Management",

ISO/IEC 10032:1995(E).

LaPlant, W. P. Jr., Lestina, G. J. Jr., Gillman, D. W., and Appel, M. V. (1996), "Proposal for a Statistical Metadata Standard", Census Annual Research Conference, Arlington, VA., March 18-21, 1996.

Lenz, H.-J. (1994), "The Conceptual Schema and External Schemata of Metadatabases", Proceedings of SSDBM-7, pp160-165, Charlottesville, VA, September 28-30, 1994.

NIST (1989), National Institute for Standards and Technology, "Information Resource Dictionary System (IRDS)", Federal Information Processing Standard (FIPS) Publication 156, April 5, 1989.

Rosen, B. and Sundgren, B. (1991), "Documentation for Reuse of Microdata from the Surveys Carried Out by Statistics Sweden", Research and Development Statistics Sweden, June 28, 1991.

StEPS (1996), "Standard Economic Processing System Document 1: Concepts and Overview", Internal Census Bureau Document, April 16, 1996.

Sumpter, R. M. (1994), "White Paper on Data Management", Lawrence Livermore National Laboratory document, 1994.

Sundgren, B. (1991a), "Towards a Unified Data and Metadata System at the Australian Bureau of Statistics - Final Report, December 2, 1991.

Sundgren, B. (1991b), "Statistical Metainformation and Metainformation Systems", R&D Report Statistics Sweden, 1991:11.

Sundgren, B. (1992), "Organizing the Metainformation Systems of a Statistical Office", R&D Report Statistics Sweden, 1992:10.

Sundgren, B. (1993), "Guidelines on the Design and Implementation of Statistical Metainformation Systems", R&D Report Statistics Sweden, 1993:4.

Sundgren, B., Gillman, D. W., Appel, M. V., and LaPlant, W. P. (1996), "Towards a Unified Data and Metadata System at the Census Bureau", Census Annual Research Conference, Arlington, VA., March 18-21, 1996.

* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997.

Appendix A: Entity Definitions for Business Data Model

Entity Name	Entity Definition	Entity Note
Data Element	A single unit of data that in a certain context is considered indivisible. It cannot be decomposed into more fundamental segments of data that have useful meanings within the scope of the enterprise.	Data is a representation of facts, concepts, or instructions in a form that allows them to be collected, organized, processed and stored in a retrievable form for communication, interpretation, or processing by human or automated means.
Emprise (Project)	An identifiable effort to generate deliverables NOT specific to a single Survey Instance	This appeared in prior models as Project
Emprise_Dataset (Project_Dataset)	A dataset containing either case level data, aggregation of case level data, or statistical manipulations of either.	This appeared in prior models as Project_dataset
Frame	A dataset containing all the cases identified for a Survey Instance based on a Survey's Universe definition	
Methodology	A structured approach to solve a problem	
Product	A finished deliverable of a Project or Survey Instance for external use.	
Program	A group of Surveys related by a common theme. A Program can be made of other Programs	
Purchaser	An external organization or individual who buys Census Bureau products	
Question	A request for one or more related pieces of information from a Case. A Question can contain other Questions	
Questionnaire	An identifiable instrument containing Questions for a particular Survey Instance	

Sample	A dataset containing a subset of a Frame for a particular set of Survey Instances, selected with a specific sampling Technique. For a census, the Sample incorporates the entire Frame.
Supplied_Dataset	A dataset acquired from sources outside the Bureau of the Census. Can be case level or aggregated/transformed data.
Supplier	An external organization which provides data to augment the Census Bureau's efforts
Survey	An investigation about the characteristics of a given Universe
Survey-Dataset	A dataset containing either case level data, aggregation of case level data, or statistical manipulations of either the case level or aggregated survey data, for a single Survey Instance
Survey-Instance	An identifiable activity which uses a System(s) to gather and process a set of Data Items from an identifiable set of cases, for a defined period of time, resulting in one or more specific deliverables
System	An identifiable process, either fully automated or computer assisted, which implements one or more Techniques to produce one or more deliverables. A System can be composed of Systems
Technique	An identifiable algorithm which is used to implement all or part of a Methodology
Universe	The total defined set of interest to one or more Surveys

Appendix B: Entity Definitions for Data Element Registry Model

Entity Name	Entity Definition
Administered Data Registration Component	A generalization for a data element, value domain, data concept, object class or property.
Classified Data Registration Component	A subtype of Administered Data Registration Component, all the data components that require classification.
Conceptual Domain	The set of possible valid values of a data element expressed without representation.
DRC Name Context	An association between an Administered Data Registration Component and a Name Context.
DRC Registration Authority	A registration authority that has registered a particular Data Registration Component.
Data Element	A single unit of data that in a certain context is considered indivisible. It cannot be decomposed into more fundamental segments of data that have useful meanings within the scope of the enterprise.
Data Element Concept	The human perception of a property of an object set, described independently of any particular representation.
Data Registration Classification Scheme	Classification schemes which are used to classify registered data.
Datatype	A category used to classify the collection of letters, digits, and/or symbols to depict values of a data element based upon the operations that may be performed on the data element.
Derivation Type	An entity used to define different types of derivations. Used to normalize the

	derivation type attribute associated for derived Data Elements and Data Element Concepts.
Derived DEC to Derived DE Mapping	An association that tracks a derivation mapping at the conceptual level to a derivation mapping at the data element level if such a mapping were to exist. This is not REQUIRED for all derivation mappings.
Enumerated VD	A list of all permissible values.
Formula	An entity that represents an algorithm to compute values. Formulas involve input quantities (Data Elements) and produce output quantities (Data Elements).
Keyword	An entity that expresses potential search keywords that users of the registry will use to search for and access Data Element Concepts.
Name Context	The system, database, standard document, or other environment in which the logical metadata class functions and the name has meaning.
Non Enumerated VD	A range used for specifying the lower limit and the upper limit of permissible values.
Object Class	A set of concepts, abstractions, or things in the natural world that can be identified with explicit boundaries and meaning and whose properties and behavior all follow the same rules.
Organization	An accredited agency authorized to declare logical metadata classes as registered. (From earlier definition of Registration Authority).
Permissible Values	Allowed values in a Value Domain
Property	A classification of any feature that humans naturally use to

	distinguish one individual object from another. It is any one of the characteristics of an object class that humans use as a label, quantity or description
Registration Authority	The organization authorized to register entries in the Registry.
Representation Class	A classification of value domains based upon the type of representational form.
Synonym Lists	A relationship that captures the fact that two distinct Data Elements have different names but the same meaning (synonym).
Value Domain	A set of Permissible Values, used to represent a Data Element.
Value Meaning	Meaning associated with Permissible Values in an Enumerated Domain.

Categorizing Event Sequences Using Regular Expressions

by Lisa Sanfilippo & John Van Voorhis

Introduction

Researchers who work with large sequential datasets are often limited in the kinds of analytic strategies they can use because of the sheer size of the data. Automated techniques for analyzing sequences were developed in the 1960s by scientists studying DNA, RNA, and proteins. In a classic volume on sequence analysis, Sankoff and Kruskal (1983) demonstrated its potential application for subjects as diverse as bird songs and macromolecules. In other work, Andrew Abbott developed "Optimal Matching" for sequence analysis in the field of sociology.

In this paper, we describe a technique for analyzing sequences using Regular Expression Matching (REM). This technique allows researchers to examine patterns in longitudinal data by condensing sequences of events into smaller, more tractable units. We also briefly discuss the development of a database structure that facilitates this kind of analysis.

Although all sequence analyses compare linear arrangements of symbols, whether in human behavior or DNA, they differ in their assumptions about what makes two sequences similar or different. Sequences in their original form often contain too much detail for useful comparison, since the possible permutations of occurrences can be limitless. Therefore, in all cases researchers must create the rules that define sequence similarity for their analyses.

Methods for determining sequence similarity are often referred to as sequence-matching algorithms. These algorithms are mathematical, and compare sequences without reference to the semantic or theoretical structures that created them. When using such methods, researchers who wish to place their analyses in an appropriate context must carefully define what events represent the phenomena of interest.

The technique described in this paper was developed as an alternative to existing algorithms and allows researchers to identify sub-patterns of events within sequences at the start of their analysis, based on theoretical or practical considerations. Because this technique operates on a single sequence at a time, it is faster than processes that require comparing many sequences to one another.

The Project

To illustrate REM, we will describe how we used it to analyze the sequence of events that led to a child's placement into foster care in three states: Illinois, Michigan, and Missouri. We were looking for systematic demographic and geographic differences among children that correlated with the events they experienced in the child welfare system. REM was developed to describe and compare the pathways the children took through this system. The data were derived from the administrative data systems of the Illinois Department of Children and Family Services, the Michigan Family Independence Agency, and the Missouri Department of Social Services.

Preliminary Data Processing

We received two data extracts from each state: one covering investigations of child abuse and neglect in the Child Protection System (CPS), and the other, services such as foster care to children in the Child Welfare System (CWS).[1]

We began by creating a project database for each state with the same essential structure. Each state's database contained tables for CPS and CWS data and one table for demographic information on the children. Next, we created an event table that contained all of the administrative events for all of the children in each system. We then transformed each child's events into a sequence variable or "history." Finally, we used regular expression matching to formulate "careers" by reducing the history sequences. At each step in the process, we preserved enough information from the previous step to retain flexibility in the subsequent steps. As the categories became broader at each step, the comparability of the data across states increased.

Creating the Event Table
In this analysis we focused on four key administrative events:

(1) indicated investigation, an investigation in which credible evidence of abuse/neglect was found,

(2) unfounded investigation, an investigation in which no credible evidence of abuse/neglect was found,

(3) case opening, when a case was opened for child welfare services, and

(4) placement, when a child was placed in a foster home or institution.

We created one record for every event a child experienced in either the CPS or the CWS. We then coded every record with a number denoting a particular event type (See "Event Codes" in Table 1). These records contained the child's ID, an event date, and an event type code (See Table 2).

Creating the History Sequences
We transformed each child's event records into a single sequence of codes, since as separate records the table structure was not appropriate for sequence analysis. To make the programming and its interpretation easier, we used only single-character codes in the history sequence. Although each history code represented a single event, a given code value could represent more than one type of event (See "History Codes" in Table 1).

We first reviewed a frequency distribution of the history sequences to identify the most common sequences and to see the repetition of patterns within and among sequences. This review also revealed data entry errors that we could correct or eliminate, such as children receiving services before their birth or children being born multiple times.

Although we had anticipated that the variation in the patterns between sequences would make them unsuitable for analyses in their present form, we had not foreseen the amount of variation in the length of the sequences. For example, examining the distribution of event sequences revealed that many children experienced only one event, while others experienced up to fifty. This wide variation in length made it difficult to make meaningful comparisons among cases and suggested that we needed a method that would not rely solely on whole-sequence comparison. Therefore, we focused our attention on identifying the sub-patterns which we had observed in the sequences.

Creating the Career Sequences

One goal of our research was to elucidate the connection between CPS investigations and a child's subsequent placement in foster care. We had three initial questions: (1) What sequences of investigations **never** resulted in a child welfare case opening and placement? (2) What sequences of investigations resulted in the child's first placement? and (3) What sequences of events resulted in the child entering the system without an investigation?

Because of our extensive work with the Illinois data and our contact with all three states regarding current and past practices and policies, we had some knowledge of what the most common patterns of events might be.

The following examples illustrate how this prior knowledge provided us with clues about what patterns to focus our attention on:

- We understood that the number of investigations a child experienced was not a critical factor in the caseworker's decision to place the child in foster care. We knew that children with histories composed solely of unfounded investigations were almost never provided with services, despite repeated contact with the department. Therefore, we believed that the number of *indicated* investigations would predict placement better than the raw number of investigations.

- We knew that, in one state, caseworkers were reluctant to remove children from their homes after only one indicated investigation unless they were in imminent danger. Thus, we expected that a child with one indicated investigation would be less likely to be placed into foster care than a child who had two or more indicated investigations.

- In all three states, we knew it was possible for children to experience a case opening and placement without an investigation of abuse or neglect, but we had no information on the frequency of such occurrences.

- Our prior analyses of the foster care data indicated that once in foster care, a child could move between placements numerous times before being returned home. Although the placements could be of different types, the child was still living away from his or her parents. As a result, we chose to treat a series of placements without a return home as one career event.

Regular Expression Matching

It became apparent in looking at the sub-patterns that they could be represented by regular expressions, a notation used widely in the computer science field for specifying and matching sequences.[2] (See Appendix.)

We created a file listing the regular expression patterns we had decided to analyze along with a "career" code for each pattern which is shown in Table 4. We grouped the patterns in passes because we knew that certain patterns occurred only at the very beginning of the history and we needed to control the generation of the matching program. The first pass was used to

remove any events that occurred before a child was born. Since we were especially interested in the first series of investigations, we created a pass that only matched to initial investigation sub-sequences. The last pass, which was applied repeatedly until the history was exhausted, contained all of the sub-patterns we were investigating. From this pattern file we generated a series of programs to transform the data.

We used the AWK programming language for both our program generator and the matching programs themselves. An AWK program is composed of a series of pattern and action pairs. It automatically reads through data files one line at a time, and each line is matched against the patterns in the order they are listed in the program. When a line contains data that matches one of the patterns, the action associated with that pattern is executed. The patterns may contain regular expressions, while the actions are written in a language similar to the C programming language.

In our project, the program generator read the pattern file containing the sub-patterns of interest to us and generated a series of programs that used those regular expression patterns to process the history data. Each program in the series corresponded to a particular pass in the pattern file. If a pattern matched to the beginning of a history sequence, the matching characters were removed and the career code for that pattern was appended to the career sequence. The child's id, history, and career were then passed to the next program for the next pass. The final program passed the data back to itself until the history sequence was empty or until a fixed number of passes had been run. If the history sequence was completely matched, a lower case 'x' was appended to the career to indicate completion. An upper case 'X' was appended if more history remained after the maximum pass limit had been reached.

Analyzing the Career Sequences

Since our analysis was limited to examining the sub-patterns that led to a child's first placement, we did not analyze children's entire careers. Instead, we only analyzed the first four career events after a child's birth.

Because the REM approach simply recoded the original history sequences, it preserved the unit of analysis, thus allowing us to attach explanatory variables such as year of first entry into the system, sex, race, and region[3]. Once this information was stored in one file, we aggregated the data by creating a crosstabulation which contained frequencies for every combination of the career sequences and the explanatory variables. These files were relatively small (fewer than 1,000 records) allowing us to import them into a spreadsheet program for final analysis and presentation.

Conclusion

The REM technique described in this paper departs from more common pattern matching methods in that it incorporates theory and practice into the actual matching process. Using this technique, researchers can test their assumptions about the structure of a sequence. It is an iterative technique that allows the analyst to explore patterns in the data and to compare them across populations simply and quickly. Because the process of developing the career file is split into several steps (i.e., creating the event table, creating the history sequences, and pattern matching), it provides many opportunities to check the data and to ensure that the processes are transforming the data correctly.

REM allows the researcher to take a very large dataset and to represent it in a much smaller form, while maintaining the critical details of event order and sequence. For example, in our Illinois database we began with an event file of over 5 million records. Transforming this file into history sequences, career sequences, and finally into a crosstabulation, decreased the size of the file by a factor of 5,000, making it significantly easier to work with.

The REM technique, as written in AWK, can save the researcher hours of processing time, in large part due to: 1) the way AWK reads data files (i.e., it automatically reads a file one record at a time) and 2) the minimal programming it requires. Performing the same analyses using a statistical software package would have required much more extensive programming and perhaps more important, would have restricted the kinds of questions we could have asked in exploring the original data.

Future Directions

Clearly, REM has a much wider application than what we have illustrated with our project. Our analysis did not utilize REM to its fullest potential. For example, instead of analyzing just the initial sequence of sub-patterns, REM could be used to analyze full careers. We could run a similar process against the career sequences to further shrink the number of categories.

Finally, we did not explore the sub-patterns in as much detail as we could have. For example, we included specific placement event types in our event table and history sequences but did not treat them as separate types. In the future, we can easily compare differences in children's histories following specific types of substitute care placements (e.g., home of a relative, private foster home, group home, etc.) based on this project's current database.

APPENDIX

Regular Expressions

In general, a character in an AWK regular expression matches itself. Some characters with special meanings in our pattern file are listed below along with some examples of their use. See the references for more details.

Special Characters Used in Regular Expressions:

Regular Expression Examples:

REFERENCES

Abbott, Andrew. 1995. Sequence Analysis. *Annual Review of Sociology*, 21:93-113.

Abbott, Andrew & Alexandra Hrycak. 1990. "Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers." *American Journal of Sociology*, 96(1): 144-185.

Abbott, Andrew & John Forrest. 1986. "Optimal Matching Methods for Historical Sequences." *Journal of Interdisciplinary History*, 16(3): 471-494.

Aho, Alfred V., Brian W. Kernighan, & Peter J. Weinberger. 1988. *The AWK Programming Language*. Reading: Addison-Wesley.

Aho, Alfred V., Jeffrey D. Ullman. 1979. *Principles of Compiler Design*. Reading: Addison-Wesley.

Forrest, John & Andrew Abbott. 1990. "The optimal Matching Method for Anthropological Data: An Introduction and Reliability Analysis." *Journal of Quantitative Anthropology* 2:151-170.

Friedl, Jeffrey E. F. 1997. *Mastering Regular Expressions*. Sebastopol: O'Reilly & Associates, Inc.

Sankoff, David & Joseph B. Kruskal eds. 1983. *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley.

NOTES

[1]. Child Protection Systems: in Illinois, the Child Abuse and Neglect Tracking System; in Michigan, the Protective Services Management Information System; and in Missouri, the Child Abuse and Neglect Data System.

Child Welfare Services Systems: in Illinois, the Child and Youth Centered Information System; in Michigan, the Children's Services Management Information System; and in Missouri, the Alternative Care Tracking System.

[2]. Regular expressions (REs) can recognize patterns which are left linear. Patterns, such as a balanced sequence of parentheses, cannot be recognized by REs because such patterns require "going-backwards" or maintaining information outside of the RE. For further information see Aho, Kernighan, and Weinberger 1988 in the references.

[3]. In all three states we differentiated the major urban area from the balance of the state.

Paper preseneted at the IASSIST/IFDO 1997 Annual Conference Odense, Denmark May 7, 1997

An Early Perspective on the "Electronic Freedom of Information Act Amendments of 1996"

*by Margaret O. Adams**

The Electronic Freedom of Information Amendments of 1996 [1]

In September, 1996, the U.S. House of Representatives and the U.S. Senate passed the "Electronic Freedom of Information Act Amendments of 1996" on a voice vote, with no debate, and with support from both Republicans and Democrats. President Clinton signed the E-FOIA bill into law in early October. The E-FOIA amendments revise the text of the Freedom of Information Act (5 U.S.C., sec. 552) or, as it is commonly known, "the FOIA," by addressing the subject of electronic records for the first time.

Many of the amendments took effect after a 180-day period, on March 31, 1997. Others do not take effect until November 1, 1997, and still others at a later date. While the FOIA and its E-FOIA amendments pertain solely to U.S. federal government records, the influence within the U.S. of this law is such that the ramifications of these new amendments can be expected to be closely watched among electronic records creators, providers, and users in both the public and private sectors of U.S. society, and potentially elsewhere.

The purpose of the Freedom of Information Act, as enacted in 1966 and amended subsequently in 1974 and 1986, is to "require agencies of the [U.S.] Federal Government to make certain agency information available for public inspection and copying and to establish and enable enforcement of the right of any person to obtain access to the records of such agencies, subject to statutory exemptions, for any public or private purpose." When he signed the E-FOIA amendments into law, President Clinton noted the important role FOIA had played in the previous 30 years in establishing an effective legal right of access to government information. He underscored the crucial need in a democracy for open access to government information by citizens. He offered his hope that as the Government uses electronic technology to disseminate more information, there will be less need for citizens to use FOIA to obtain government information.

The legislative history prepared by the [U.S.] House Committee on Government Reform and Oversight as background for the E-FOIA amendments identifies the purpose of the amendments as providing for "public access to information in an electronic format, and for other purposes..." Their history highlights several key aspects of the E-FOIA amendments, from the legislators' perspective:

-- electronic records: the amendments make explicit that electronic records are subject to the FOIA. Furthermore they acknowledge the increase in the Government's use of computers and encourage federal agencies to use computer technology to enhance public access to Government information.

— format request: with implementation of the E-FOIA amendments, requestors may request records in any form or format in which an agency maintains the records. Also, agencies must make a "reasonable effort" to comply with requests to furnish records in [any] formats specified by the requestor, "if the record is readily reproducible by the agency in that form or format." This change reverses a legal opinion that dates from 1984 and which formed the basis for federal agency practices since that time. This change will be discussed further, below.

-- redaction: agencies redacting electronic records (deleting part or parts of an electronic record [or an electronic records file] to prevent disclosure of material that is exempted from release), shall indicate the amount of information deleted on the released portion of the record, unless doing this would harm an interest protected by the exemption. Further, "if technically feasible, the amount of the information deleted shall be indicated at the place in the record where such deletion is made."

— expedited processing: the amendments establish that certain categories of requestors would receive priority treatment of their requests if failure to obtain information in a timely manner would pose a significant harm. The first such category are those who might reasonably expect that delay in obtaining the information could pose an imminent threat to life or physical safety of an individual. The second category includes requests made by a person(s) primarily engaged in the dissemination of information to the public, e.g., the media, and involving a compelling urgency to inform the public.

-- multitrack processing: the writers of the amendments

created an incentive for requestors to submit narrowly specified information requests to federal agencies by allowing agencies to establish procedures to process FOIA requests of various sizes on different tracks rather than on a first-received, first-responded-to order. The assumption here is that requests for specific or smaller amounts of information can be completed quickly, so responses to such requests no longer need to be in a queue with more general or larger-volume requests.

-- agency backlogs: in an effort to ameliorate the phenomenon of significant backlogs in responding to FOIA requests in many federal agencies, the amendments stipulate that agencies can no longer delay responding to FOIA requests because of "exceptional circumstances" if such circumstances simply result from a predictable agency request workload.

— deadlines: the amendments extend the deadline for agencies to respond with an initial determination to a FOIA request to 20 working days, from the previous deadline of 10 working days.

From the perspective of the executive branch of the government, the part of the federal government that has to implement the E-FOIA amendments, the effects of this bill are highlighted somewhat differently. According to the Justice Department's newsletter, FOIA Update, a major change of the amendments concerns the maintenance of electronic access in agency reading rooms. Prior to the E-FOIA amendments, agencies were required to make three categories of records routinely available for public inspection and copying: final opinions rendered in the adjudication of administrative cases, specific agency policy statements, and administrative staff manuals that affect the public. The amendments add to the categories of reading room records and also establish a requirement for electronic availability of reading room records.

The new category of records that agencies have to make available in their reading rooms as of March 31, 1997, includes any records processed and disclosed in response to a FOIA request that "the agency determines have become or are likely to become the subject of subsequent requests for substantially the same records." By the eve of the new century, December 31, 1999, agencies are to have a general index to FOIA-released records and are to make this index available by computer telecommunications.

Theoretically the idea is that making records in greatest demand accessible in an agency reading room should satisfy most future demand for those records. But, even for federal agencies that maintain public reading rooms in their regional offices throughout the country, this expectation may not be met. It suggests that the bill drafters assume that most of the public's demand for records under FOIA can be satisfied by having an agency reading room where researchers can come to "read" records. Yet the amendments stipulate that anytime an agency receives a FOIA request for records, the agency must treat the request in the standard FOIA fashion, regardless of whether it also makes these records available in its reading room or online. In other words, even though it already makes such records available in an agency reading room or online, it must respond formally to the requestor within a 20-day time period, and provide the records under whatever guidelines and fees it has established for processing requests under FOIA. Note here: a FOIA request is any request for records that invokes or mentions the FOIA, or Freedom of Information Act.

The E-FOIA amendments, as suggested above, also expand the concept of an agency reading room to what some are referring to as "electronic" or virtual reading rooms. The amendments require that agencies use electronic information technology to enhance the availability of their reading room records. And, for any "newly created reading room records [i.e., records created on or after November 1, 1996 that are in the category of "reading room records"], agencies must, by November 1, 1997, make these records available to the public by electronic means. Preferably this new electronic availability should be via computer telecommunications, i.e., in the form of on-line access, such as from a World Wide Web site(s) established to serve as "electronic reading room(s)." While the amendments do not explicitly state that agencies are to continue to maintain their conventional reading rooms, the advice in the Justice Department's FOIA Update, is that they are to do this.

In other words, the three categories of administrative and policy records traditionally maintained by agencies in their public reading rooms, plus any records released under FOIA that the agency determines are likely to become the subject of subsequent requests, must be made available for public inspection and copying in the agency public reading room. Further, any of these reading room materials created after November 1, 1996 must also be made available to the public by computer telecommunications by November 1, 1997.

Finally, there are two additional new requirements that may have significant implications for agencies seeking to comply with both the spirit and the letter of the E-FOIA amendments. The first has already been mentioned earlier: agencies are to make records available in any form or format requested by the person if the record(s) is(are) readily reproducible by the agency in that form or format. The second new requirement was somehow not highlighted in the legislative history. Yet compliance with it may require substantial reorientation in the way federal agencies treat FOIA requests for information in federal records, when those records are in an electronic format. This is the requirement that states: "an agency shall make reasonable efforts to search for the records [responsive to a request]

except when such efforts would significantly interfere with the operation of the agency's automated information system." "Search" is defined as "to review, manually or by automated means, agency records for the purpose of locating those records which are responsive to a request."

Clearly, the impetus for the E-FOIA amendments just described comes from evolution in the use of electronic computer technology by both government agencies and the federal records-seeking public. It also reflects growing expectations for public access to more and more government information that has accompanied the proliferation of computer technology, especially personal computers. Some aspects of the amendments, however, reflect more than natural evolutionary change. The new right for requestors to choose the format in which they expect to receive federal records reverses long-standing legal opinion. The requirement that agencies use computer technology to search for records in electronic form reverses widespread federal practice rooted in a series of court rulings. To understand how or why these changes came to be law, it may be helpful to consider the historical context from which they emerged.

"The Freedom of Information Act in the Information Age..."

Our colleague Tom Brown recently published an article that examines the historical context of the FOIA. He discusses the case law on how FOIA related to computerized records prior to enactment of the E-FOIA amendments.[2] He notes that FOIA guarantees any person the right to gain access to records unless the records contain information on matters specifically excluded under one or more of the nine exemptions identified in the FOIA. Further if a portion of a record is exempt from disclosure then a reasonably segregable portion of the record is to be provided after deletion of the portions which are exempt. Brown notes that in interpreting the FOIA statute, the courts have consistently ruled that agencies are not required "to create records in order to respond positively to a FOIA request," i.e., to provide records in response to a FOIA or to segregate exempted portions of records in order to release them in response to a request.

Even prior to the E-FOIA amendments, the courts seemed to have resolved the question of whether electronic materials were subject to the FOIA. In a 1982 case cited by Brown, a federal appeals court ruled that "[C]omputer-stored records, whether...in the central processing unit, on magnetic tape or in some other form, are still records for the purposes of the FOIA." Further, a 1989 Department of Justice survey of federal agency practices found that government-wide practice also affirmed that electronic records may be records under the FOIA.

The question of whether the FOIA required federal agencies to provide requestors with records in the formats of the requestors choosing had generally been decided in ways that allowed the federal agencies to make that choice. For example, Brown discusses a 1984 case, Dismukes v Department of the Interior. The case centered on a FOIA request to Interior's Bureau of Land Management (BLM) for a list of participants in oil and gas leasing lotteries in California. Dismukes, on behalf of the National Wildlife Federation — a private organization, had filed a FOIA for these records and requested that they be provided on 9-track, 1600 bpi magnetic tape, in an IBM-compatible format, and with file dumps and file layouts. BLM's Office of Surface Mining provided computer printouts of the requested records. The National Wildlife Federation appealed this response, arguing that "it is impossible to work with such volumes of data without having it in computer form." The Court of Appeals ruled that "the computer printout was fully responsive to the...request." As a result of this ruling, the Department of Justice advised federal agencies that they, not requestors, had the right "to choose the format of disclosure, so long as the agency chooses reasonably." This same principle was upheld a few years later when a requestor appealed the response of the Central Intelligence Agency (CIA) to a FOIA request for an index of documents that the CIA had previously released. In response, the CIA provided 5000 pages of printouts, arranged by date of release of the item in the index. In its appeal, the requestor asked that this information be made available on tape or disk. The court ruled however, that the "information [the CIA had provided] was in a "reasonably accessible form" and furthermore, that the agency was not obligated to provide in electronic format, records it had already provided in paper copy." These rulings, Brown suggests, provide evidence that the courts were in denial of the computer age.

While the courts may have been in denial, or in ignorance of the computer age, Brown also makes clear that their rulings were predicated on the basic premise that agencies are not "expected to be private research firms, ...subject to every beck and call of a requester." To bolster this assessment, he cites a number of rulings. One, by a federal district court in Pennsylvania in 1988, clarified that FOIA does not require agencies to write new computer programs to search for data not already compiled for agency purposes. Another, a federal appeals court, determined that "the FOIA dictum to release reasonably segregable portions [of records] does not require creation of a ...summary file because it is not functionally analogous to manual searches" for records that contain information responsive to a request. Another concluded that the FOIA "in no way contemplates that agencies, in providing information to the public, should invest in the most sophisticated and expensive form of technology."

On the issue of format, Brown also shows that as early as 1988, the Administrative Conference of the United States (until October 1995, the federal government agency

responsible for studying and recommending improvements in administrative procedures to executive branch agencies) had recommended that in responding to FOIA requests, "agencies should provide electronic information in the form in which it is maintained or, if so requested, in such other form as can be generated directly and with reasonable effort."

Taken together, these rulings of the courts and the practices of some agencies responding to FOIA requests for information in electronic formats point to the rationale for the E-FOIA amendments. So, now, what can be their anticipated impact?

Application of the E-FOIA Amendments

In general, it is far too early to know how federal agencies will adapt their practices to be in compliance with both the spirit and the letter of the E-FOIA amendments. Similarly, it is much too soon to have any idea whether the cumulative effect of compliance with the E-FOIA amendments will result, as President Clinton said he hoped, in less need by citizens to use FOIA to obtain government information.

There are a few things that we can suggest at this time however. J. Timothy Sprehe, writing in the Federal Computer Week (January 6, 1997) suggests that the principal benefits of the law "lie in the fact that EFOIA overturns two bad court decisions." One of these was a ruling that the National Library of Medicine's on-line information systems were not agency records for the purposes of FOIA. The other was the Dismukes case discussed above, where the court had ruled that an agency had the prerogative to decide the format in which it fulfills a FOIA request. But beyond this, Sprehe does not consider the E-FOIA to be "a great leap forward," except in the sense that the "fact that it was passed at all may be an important reminder that the public has rights of access to electronic as well as paper-based information resources."

Another article, this one in a trade newspaper, Washington Technology (April 24, 1997), quotes an analyst for the Federation of American Scientists as saying "agencies must undergo a cultural transformation to accommodate [to] the requirements of [E-FOIA]...the law does not change reality...but it provides an incentive to modernize." What he is referring to was further enunciated in this same article by a Washington lawyer who raises the question of whether federal agencies have the hardware, software, and personnel that will enable them to be in compliance with the E-FOIA amendments. As he states, the amendments "raise the issue of the availability of suitable software and the equipment to handle it, such as a client-server system with sufficient storage capacity and database software with advanced features. It also raises the issue of being able to hire and find personnel who are sufficiently trained to use what is sophisticated search and retrieval software to comply with requests."

The principles that underlie the FOIA and now the E-FOIA amendments are firmly based on the principles of the U.S. Constitution and its Bill of Rights. Yet, the IASSIST community, sophisticated and knowledgeable as it is in matters regarding maintenance and access to electronic records and information, might well ponder the implications for federal government agencies seeking to be in compliance with the FOIA as amended. Keep in mind that in general, the data community that IASSIST represents, offers or supports services for researchers knowledgeable in the structure and use of electronic data. Social science researchers and those who offer support services to social scientists, such as data archivists and librarians, are among those most likely to expect increasing expansion in access to electronic government information. Yet the need by social scientists, generally, for access to administrative and programmatic databases to use as sources for rigorous research and analysis purposes, are quite different than the needs reflected in a significant portion of the requests that, for example, the Center for Electronic Records at the U.S. National Archives and Records Administration, receives.

Since our holdings reflect the records of the entire federal government, we can assume that at least in some senses, the requests we receive mirror those received by all federal agencies. In our case, for the first six months of the current fiscal year, approximately one-third of the inquiries (over 1100 — and they generated over 1800 separate responses) requested specific items of information from records in our holdings. Very few of these invoked the FOIA, an indication perhaps, of the well-known practice of the [U.S.] National Archives and Records Administration (NARA) to treat all requests as if the requestor had invoked the FOIA, thus negating the need for requestors to use it in order to receive the records or the information in records that they seek. That is, it is the policy of NARA to respond to all inquiries in a timely manner, and as responsively as possible. The extent to which it can successfully do this is in large measure a reflection of how informed the request is — how specifically it identifies the information sought, and whether the manner in which the specificity of the request reflects the description that NARA has for the relevant records.

In our particular case, most requests for specific records or for information from specific records, pertained to the casualty records from the Korean and Vietnam conflicts, two of our best-known electronic records files. We have long experience in responding to such requests, and the E-FOIA amendments will have virtually no impact on the manner in which we handle responses to these inquiries — from printouts of the full files, or by using some pc-based versions of the files, with retrieval software, that a small business vendor has created and provided to us in beta-test

format. Two individuals have developed web sites with these records where anyone can access them. Interestingly, such widespread availability of these records seems to have had no impact on the continuing demand that we receive for specific information from them. In microcosm at least, this experience suggests that the ready availability of records does not stem the direct demand for them.

So, our real challenge in complying with the E-FOIA requirement to search electronic records for information responsive to a request will not be in regard to the casualty records. Rather it will be in responding to requests for information that may reside in any of the other 30,000 (and growing) electronic records files in our holdings. And this is but a reflection of the challenges facing federal agencies as a whole. While each agency only receives requests for information from its records, whereas NARA receives requests for information from archival records of the entire federal government, the records in agencies are more current and thus they are in more urgent demand than archival records usually are. There is no arguing amongst us that technological innovation makes access to public information more efficient and varied in ways that none of us could have imagined even just a few years ago. But, we also know too that with each innovation, new complexities and possibilities have also presented us with the reality of a seemingly infinite rise in the level of expectations for what we can do or offer. Keeping pace with such expectations, or rather, determining the nature of the "reasonable" effort by which we measure the limits for responding to those expectations, is perhaps the greatest challenge for those seeking to provide service to citizens that is responsive both to the spirit and the law embodied in the FOIA, as amended. The experiences of the data community can assist and influence these determinations. Outreach by the data community to the larger universe can perhaps also help to inform their expectations.

NOTES

1. The first portion of this paper is based primarily upon the Committee on Government Reform and Oversight, U.S. House of Representatives Report 104-795, 104th Congress, 2d Session, Electronic Freedom of Information Amendments of 1996: Report [to accompany H.R. 3802], September 17, 1996. This paper also draws upon U.S. Department of Justice, Office of Information and Privacy, FOIA Update, Vol. XVII, No. 4, Fall 1996.

2. Thomas Elton Brown, "The Freedom of Information Act in the Information Age: The Electronic Challenge to the People's Right to Know," American Archivist, Vol. 58, Spring 1995, pp. 202-211. Much of the analysis in this section of this paper draws upon Brown's article.

* Presented at IASSIST-97 (May 9, 1997), Odense Denmark

IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

• • • •

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Services and Technology (IASSIST)** is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**.

Membership fees are:

Regular Membership. $40.00 per calendar year.
Student Membership: $20.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription: $70.00 per calendar year (includes one volume of the Quarterly)

Membership form

I would like to become a member of IASSIST. Please see my choice below:
Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details: http://datalib.library.ualberta.ca/iassist/mbrship2.html

- ☐ $40 (US) Regular Member
- ☐ $20 Student Member
- ☐ $70 Subscription (payment must be made in US$)
- ☐ List me in the membership directory
- ☐ Add me to the IASSIST listserv

Please make checks payable, *in US funds*, to IASSIST and Mail to:

IASSIST,
Assistant Treasurer
JoAnn Dionne
50360 Warren Road
Canton, MI 48187
USA

Name: _____
Job Title: _____
Organization: _____
Address: _____

City: _____ State/Province: _____
Postal Code: _____ Country: _____
Phone: _____ FAX: _____
E-mail: _____ URL: _____

Return Undelivered Mail To:

IASSIST QUARTERLY
c/o Wendy Treadwell
1758 Pascal St. North
Falcon Heights, MN 55113
USA

Serials Department(SERLIBS82186344), Univ of
North Carolina-Chapel Hill
Chapel Hill, NC 27514-8890
U.S.A